Edition eGov-Campus

Holger Hünemohr · Jörn von Lucke · Jürgen Stember · Maria A. Wimmer *Hrsg.*

Moreen Heine · Anna-Katharina Dhungel · Tim Schrills · Daniel Wessel

Künstliche Intelligenz in öffentlichen Verwaltungen

Grundlagen, Chancen, Herausforderungen und Einsatzszenarien







Edition eGov-Campus

Reihe herausgegeben von

Der eGov-Campus bietet rund um das Thema E-Government und Verwaltungsinformatik Bildungsangebote auf Hochschulniveau. Das breite Spektrum der eGov-Campus-Themen reicht von Prozessmanagement, Verwaltungsportalen, IT-Architektur bis hin zu Informationssicherheit und Künstliche Intelligenz in der Verwaltung. Die qualitätsgesicherten Lernmodule werden von führenden Hochschulen und Professor:innen in Deutschland angeboten. Das Ziel ist, Studierenden und Beschäftigten in der Verwaltung zum Thema Digitalisierung des öffentlichen Sektors und E-Government hochwertige Online-Bildungs- und Weiterbildungsmöglichkeiten zu bieten, um erforderliche Kompetenzen und Qualifikationen zu erlangen.

Die Bücher aus der Reihe Edition eGov-Campus begleiten die Module auf der Lernplattform www.egov-campus.org. Sie stellen die Inhalte im klassischen Buchformat in gedruckter und digitaler Form zur Verfügung und machen sie individuell und schnell erschließbar. Die Bücher werden mit digitalen Flashcards von Springer Gabler | Springer Nature erweitert, mit dem die Lernenden ihr Wissen anhand von Fragen und Antworten spielerisch überprüfen können. Die Bücher sind als Open-Access-Publikation in der eBook-Version kostenfrei zugänglich.

Der eGov-Campus wird vom IT-Planungsrat auf Initiative der Bundesregierung gefördert.

Moreen Heine ·
Anna-Katharina Dhungel · Tim Schrills ·
Daniel Wessel

Künstliche Intelligenz in öffentlichen Verwaltungen

Grundlagen, Chancen, Herausforderungen und Einsatzszenarien

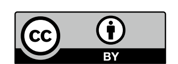


Moreen Heine
Institut für Multimediale und Interaktive Systeme
Universität zu Lübeck
Lübeck, Deutschland

Anna-Katharina Dhungel
Institut für Multimediale und Interaktive Systeme
Universität zu Lübeck
Lübeck, Deutschland

Tim Schrills
Institut für Multimediale und Interaktive Systeme
Universität zu Lübeck
Lübeck, Deutschland

Daniel Wessel
Institut für Multimediale und Interaktive Systeme
Universität zu Lübeck
Lübeck, Deutschland



ISSN 2751-7357 ISSN 2751-7365 (electronic)
Edition eGov-Campus
ISBN 978-3-658-40100-9 ISBN 978-3-658-40101-6 (eBook)
https://doi.org/10.1007/978-3-658-40101-6

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Planung/Lektorat: Rolf-Günther Hobbeling
Springer Gabler ist ein Imprint der eingetragenen Gesellschaft Springer Fachmedien Wiesbaden GmbH und ist ein Teil von Springer Nature.
Die Anschrift der Gesellschaft ist: Abraham-Lincoln-Str. 46, 65189 Wiesbaden, Germany

Ihr Bonus als Käufer dieses Buches

Als Käufer dieses Buches können Sie kostenlos unsere Flashcard-App „SN Flashcards“ mit Fragen zur Wissensüberprüfung und zum Lernen von Buchinhalten nutzen. Für die Nutzung folgen Sie bitte den folgenden Anweisungen:

1. Gehen Sie auf **https://flashcards.springernature.com/login**
2. Erstellen Sie ein Benutzerkonto, indem Sie Ihre Mailadresse angeben und ein Passwort vergeben.
3. Verwenden Sie den Link aus einem der ersten Kapitel um Zugang zu Ihrem SN Flashcards Set zu erhalten.



Ihr persönlicher SN Flashards Link befndet sich innerhalb der ersten Kapitel.

Sollte der Code fehlen oder nicht funktionieren, senden Sie uns bitte eine E-Mail mit dem Betreff **„SN Flashcards“** und dem Buchtitel an **customerservice@springernature.com.**

Vorwort

Dieses Buch basiert auf dem eGov-Campus-Kurs „KI in öffentlichen Verwaltungen“ (www.egov-campus.org). Ziel des Kurses und damit auch dieses Buches ist es, die Kompetenzentwicklung im Umgang mit KI-Anwendungen sowie zur Gestaltung von KI-Anwendungen zu fördern. Darüber hinaus bilden die im BMBF-Projekt „KIOEV — KI in öffentlichen Verwaltungen“ entwickelten Online-Micro-Lerneinheiten, die über die Lernplattform KI-Campus (www.ki-campus.org) zugänglich sind, eine weitere Grundlage für die Inhalte dieses Buches.

Wir bedanken uns herzlich für die Unterstützung bei der Erstellung dieses Buches bei Eva Beute, Marvin Sieger, Bastian Mannerow, Eric Förster und Carola Mohrmann sowie bei allen Beteiligten und Gesprächspartnern der Projekte zur Erstellung der beiden Online-Kurse.

Vielen Dank insbesondere an das gesamte eGov-Campus-Team und das Team der Springer Fachmedien Wiesbaden GmbH.

Dieses Buch wurde um digitale Lernkarten (Flashcards) erweitert. Diese ermöglichen es dem Leser, das Gelernte zu überprüfen und das Wissen aus dem Buch zu vertiefen. Bitte laden Sie die kostenlose Flashcards-App des Springer-Verlags aus dem Google oder Apple-Store auf Ihr Handy oder Tablet. Den Zugangscode zu den Flashcards für dieses Lehrbuch finden Sie unten.

Lübeck
April 2023

Moreen Heine
Anna-Katharina Dhungel
Tim Schrills
Daniel Wessel

Inhaltsverzeichnis

Über die Autoren

Anna-Katharina Dhungel arbeitet als wissenschaftliche Mitarbeiterin am Institut für Multimediale und Interaktive Systeme im Bereich E-Government und Open Data Ecosystems. In ihrer Forschung untersucht sie unterschiedliche Einsatzszenarien von Künstlicher Intelligenz im Öffentlichen Sektor, insbesondere in der Judikative.

Moreen Heine ist Professorin für E-Government und Open Data Ecosystems an der Universität zu Lübeck (Institut für Multimediale und Interaktive Systeme) und wissenschaftliche Leiterin des Joint eGov and Open Data Innovation Labs. Sie forscht zu menschzentrierten und prozessorientierten Anwendungen im öffentlichen Sektor.

Tim Schrills arbeitet am Institut für Multimediale und Interaktive Systeme in Lübeck. Er fokussiert sich in seiner Forschung auf die Interaktion zwischen Menschen und intelligenten Systemen und untersucht, wie diese Systeme vertrauenswürdig und nachvollziehbar gestaltet werden können.

Daniel Wessel ist Postdoc am Institut für Multimediale und Interaktive Systeme (IMIS) an der Universität zu Lübeck. Er arbeitet als wissenschaftlicher Mitarbeiter im Schnittbereich von Psychologie und Technologie in Forschung und Lehre, insbesondere mit den Themen EGovernment und Evaluationen.

Abbildungsverzeichnis



1 Einleitung

Zusammenfassung

Dieses einleitende Kapitel gibt einen Überblick über die Inhalte des Buches (1.1). Es klärt begriffliche Grundlagen in aller Kürze (1.2) und zeigt die Anwendungsbereiche für KI in öffentlichen Verwaltungen auf (1.3).

Flashcards zur Lernunterstützung

Mit der kostenlosen Flashcard-App „SN Flashcards" können Sie Ihr Wissen anhand von Fragen überprüfen und Themen vertiefen.

Für die Nutzung folgen Sie bitte den folgenden Anweisungen:

1. Gehen Sie auf https://flashcards.springernature.com/login
2. Erstellen Sie ein Benutzerkonto, indem Sie Ihre Mailadresse angeben und ein Passwort vergeben.
3. Verwenden Sie den folgenden Link, um Zugang zu Ihrem SN Flashcards Set zu erhalten: https://sn.pub/kCm6Vh

Sollte der Link fehlen oder nicht funktionieren, senden Sie uns bitte eine E-Mail mit dem Betreff „SN Flashcards" und dem Buchtitel an customerservice@springernature.com.

M. Heine et al., *Künstliche Intelligenz in öffentlichen Verwaltungen*, Edition eGov-Campus, https://doi.org/10.1007/978-3-658-40101-6_1

1.1 Überblick über die Themen des Buches

Künstliche Intelligenz gewinnt auch im öffentlichen Sektor zunehmend an Bedeutung. Chatbots, Systeme zur Klassifizierung von Dokumenten oder zur Erkennung von Schäden, zum Beispiel an Straßen, sind bereits im Einsatz. Es stellen sich einige Fragen, die dieses Buch aufgreift: Was bedeutet Künstliche Intelligenz? Wie können KI-Systeme im öffentlichen Sektor genutzt werden? Dies wird anhand verschiedener Einsatzgebiete beispielhaft erläutert. Auch grundsätzliche Fragen sind zu beleuchten: Welche Erwartungen und Ziele werden mit dem KI-Einsatz verbunden? Welche Probleme werden adressiert? Dies sind Fragen der strategischen Ausrichtung. Auch Aspekte der Governance, also Steuerungsfragen spielen eine Rolle. Außerdem werden grundsätzliche Herausforderungen und Grenzen diskutiert. Ein besonderer Schwerpunkt liegt auf der Betrachtung der Beziehung und der Interaktion zwischen Mensch und KI-System. Dabei wird auch die Erklärbarkeit der Funktionsweise eines KI-Systems aufgegriffen. Nach dem Lesen dieses Buches sind Sie in der Lage

- KI-Anwendungsfälle und Potenziale im öffentlichen Sektor zu identifizieren,
- KI-Methoden im Überblick zu verstehen,
- Grenzen und Herausforderungen bei der Anwendung im öffentlichen Sektor zu diskutieren und
- Anforderungen an KI-Anwendungen im öffentlichen Sektor zu erarbeiten.

Kenntnisse im Bereich der Informatik werden nicht vorausgesetzt.

1.2 Zum Begriff

Künstliche Intelligenz ist ein schwieriger Begriff. Das beginnt mit der Frage, was genau eigentlich Intelligenz ist. Wie können wir Intelligenz beschreiben oder sogar messen und sollten wir wirklich versuchen, sie künstlich zu schaffen? Ursprünglich ging es um das Ziel, Maschinen zu entwickeln, die sich so verhalten, als wären sie intelligent. Problematisch ist, dass recht simple Mechanismen dazu führen können, dass wir Maschinen intelligent finden. Das ist eine Frage der Erwartungshaltung und der bereits gemachten Erfahrungen mit Maschinen. Stellen wir uns einen Zeitreisenden vor, der aus dem 19. Jahrhundert in die 70er Jahre reist und dort einen Taschenrechner vorfindet. Wird diese Person den Taschenrechner als intelligent empfinden? Und wenn die Person weiterreist und beobachtet, wie ein Kind einen Sprachassistenten nach dem Wetter

fragt und eine angemessene Antwort erhält, wird die Person denken, dass dieser Sprachassistent über Intelligenz verfügt? Es ist daher wichtig, einen Bezug herzustellen. Das gelingt über einen Vergleich zu aktuellen menschlichen Fähigkeiten. KI befasst sich also mit der Frage, wie Computer Aufgaben bewältigen können, bei denen Menschen derzeit überlegen sind (vgl. Rich, 1985). Diese Formulierung macht deutlich, dass der Begriff KI hier keine spezifische Technologie oder eine bestimmtes System meint. Künstliche Intelligenz ist eine Teildisziplin der Informatik und nutzt Methoden und Erkenntnisse aus verschiedenen Gebieten, zum Beispiel Logik, Statistik, Bildverarbeitung, Linguistik, Psychologie und Neurobiologie. Was ist nun heute besonders relevant? Bei welchen Aufgaben sind Menschen heute klar überlegen? Menschen sind besonders gut darin, mit neuen Situationen und neuen Anforderungen umzugehen. Im Fokus sind demnach Verfahren, die zur Entwicklung von KI-Systemen führen, die ebenfalls in der Lage sind, möglichst schnell und angemessen zu reagieren. Daher ist maschinelles Lernen aus heutiger Sicht ein zentrales Teilgebiet der KI (Ertel, 2016, S. 3).

1.3 Anwendungsoptionen für KI im öffentlichen Sektor

Wo können wir KI nun konkret anwenden? Die Handlungsfelder und Aufgabengebiete öffentlicher Verwaltungen sind vielfältig. Daher ist es hilfreich, sich dieser Frage über grundsätzliche Einsatzgebiete zu nähern. Unterschieden werden im Folgenden die drei Bereiche *Informieren und Kommunizieren*, *Erkennen* und *Agieren* (siehe Abb. 1.1).

KI-basierte Systeme können zum einen genutzt werden, um zu informieren und zu kommunizieren, zum Beispiel über Chatbots und Sprachassistenten. Auch bei erhöhtem Bedarf ist es so möglich, 24/7 erreichbar zu sein. Beschäftigte im öffentlichen Sektor werden von einfachen Beratungsgesprächen entlastet. Ein weiterer Vorteil liegt in der Möglichkeit, mehrsprachige Kommunikationsangebote zu schaffen.

Außerdem können KI-basierte Systeme helfen, Auffälligkeiten zu erkennen, wie beispielsweise Anomalien, die auf Fehler oder Betrug hinweisen könnten. Auch die Identifizierung von Personen auf Basis von Bild- und Tonmaterial ist möglich, ebenso die Analyse der Echtheit von Nachweisen und Urkunden, wie zum Beispiel Studienbescheinigungen. Dabei ist es auch erforderlich, die Relevanz für die aktuelle Entscheidung festzustellen. Handelt es sich also um das richtige Semester und eine anerkannte Hochschule und ist das Dokument echt? Ein interessantes Anwendungsgebiet ist die Klassifizierung von Dokumenten.

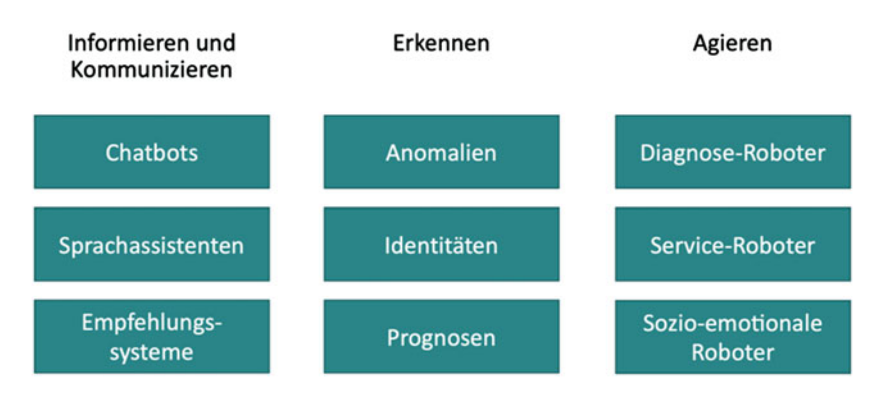


Abb. 1.1 Anwendungsbereiche von KI im öffentlichen Sektor

Für komplexe Planungsvorhaben können dann relevante Dokumente, zum Beispiel Grundbuchauszüge, automatisiert identifiziert werden. Im nächsten Schritt können auch die relevanten Daten, etwa zu Grundbesitzern, die in den Prozess einzubeziehen sind, automatisiert ausgelesen werden. Hierzu stehen Lösungen am Markt zur Verfügung. Ein weiteren Feld umfasst Anwendungen, die Prognosen über die Zukunft ermöglichen. Dies kann zum Beispiel Steuerungs- und Managemententscheidungen verbessern.

Schließlich können KI-basierte Systeme auch agieren. Diagnose- und Service-Roboter sind hier denkbar. Zum Beispiel zur Erfassung der Zustände von Bauwerken, wie Brücken oder Straßen, aber auch zu Einsätzen in einem gefährlichen Umfeld, beispielsweise nach einem Brand zur Untersuchung der Einsturzgefährdung eines Bauwerks. Sozio-emotionale Roboter könnten in Empfangssituationen oder als Begleitung, beispielsweise auf einem unübersichtlichen Verwaltungscampus, genutzt werden. Roboter können auch Ordnungswidrigkeiten und unerwünschtes Verhalten im öffentlichen Raum erfassen. Der Roboter ermahnt dann die jeweiligen Personen und kann auch menschliche Ordnungshüter zu Hilfe rufen. Das sind mehr oder weniger visionäre Ideen, vieles finden wir jedoch auch schon in der Realität vor. Eine Herausforderung ist, die konkreten Einsatzszenarien zu identifizieren, bei denen die Nutzung von KI-Systemen vielversprechend ist.

1.4 Aufgaben zum eigenen Anwendungsfall

Sie haben jetzt einen Überblick über mögliche Anwendungsoptionen von KI-Systemen im öffentlichen Sektor.

- Wählen Sie vor diesem Hintergrund einen geeigneten Anwendungsbereich für eine konkrete Organisation aus. Diese Organisation können Sie selbst wählen.
- Welche Ziele verfolgen Sie mit Ihrer KI-Anwendungsidee? Begründen Sie die Auswahl des Anwendungsbereichs.
- Skizzieren Sie nun ein konkretes Projekt.

Dieses Projekt bearbeiten Sie nach jedem Kapitel weiter. Nach den Grundlageneinheiten sollten Sie überprüfen, inwiefern Ihr KI-Projekt geeignet ist und gegebenenfalls nochmal anpassen. Daher bearbeiten Sie die ersten Aufgaben zum eigenen Anwendungsfall bis Kap. 5 (KI-Strategie) zunächst vorläufig.

Literatur

Ertel, W. (2016). *Grundkurs Künstliche Intelligenz* (Bd. 4). Springer Fachmedien Wiesbaden.

Rich, E. (1985). Artificial intelligence and the humanities. *Computers and the Humanities 19,* 117–122.

2 Grundlagen: Input

Zusammenfassung

Häufig wird davon gesprochen, dass Daten wertvoll sind. Daten würden Entscheidungen unterstützen, Prozesse beschleunigen oder Erkenntnisse liefern. Diese Aussagen sind so allerdings nicht korrekt, denn Daten allein haben keinen intrinsischen Mehrwert, sie sind zunächst einfach nur „da". Erst eine Verwertung und Verarbeitung von Daten führt zu einem Wertzuwachs. Dennoch ist es wichtig, dass vor der Verarbeitung ein präzises Verständnis darüber vorhanden ist, welche Art von Daten vorliegen. Nur wenn man dieses tiefreichende Verständnis von Daten hat, kann man diese anmessen nutzen und den angesprochenen Mehrwert erzeugen. Daher wird in dieser Lerneinheit der Fokus darauf liegen, Daten zu verstehen (2.2), die Qualität von Daten zu beurteilen (2.4), die Beziehungen zwischen Daten zu analysieren (2.5) sowie mögliche Verzerrungen von Daten zu erkennen (2.6). Das angeeignete Wissen kann im eigenen Anwendungsfall eingesetzt werden (2.7).

2.1 Einleitung

In dieser Lerneinheit steht die Voraussetzung für das Trainieren eines KI-Systems im Fokus – womit muss das System sozusagen „gefüttert" werden bzw. Was ist der Input? Die Antwort hierauf erscheint simpel: ein KI-System benötigt Daten. Doch Daten sind nicht gleich Daten, eine Excel-Tabelle unterscheidet sich von einem Foto, eine Audio-Aufnahme ist etwas anderes als ein handschriftlich ausgefülltes Formular. Bevor eine präziere Betrachtung der Funktionsweisen von KI-Systemen durchgeführt werden kann, muss daher vorab eine Analyse der vorliegenden Daten stattfinden.

M. Heine et al., *Künstliche Intelligenz in öffentlichen Verwaltungen*, Edition eGov-Campus, https://doi.org/10.1007/978-3-658-40101-6_2

Übung

Für einen Einstieg in das Thema soll zunächst ein bestimmtes Formular aus der öffentlichen Verwaltung näher betrachet werden. Es handelt sich hierbei um den Antrag auf eine Beihilfe für Renovierungskosten gemäß § 22 Absatz 1 SGB II, der Antrag ist über das Jobcenter Kreis Warendorf verfügbar. Die Inhalte der meisten Felder sind konkret vorgegeben, zum Beispiel der Name oder das Einzugsdatum in die Wohnung (Abb. 2.1).

Das Formular enthält jedoch auch eine indirekte „Warum"-Frage: „Begründung für die Beantragung der Renovierungskostenbeihilfe" (Abb. 2.2).

Ein solches Freitextfeld ist für Computer in der Regel schwierig zu verarbeiten. Bei der Angabe des Namens kann man etwa festlegen, dass in dem Feld keine Sonderzeichen oder Ziffern auftreten dürfen. Dadurch werden die eingegebenen Daten gleichartig und sind leichter auszuwerten. In einem Freitextfeld kann es jedoch notwendig sein, Ziffern zu nutzen, wie hier etwa „In der Wohnung sind 3 Heizungen defekt, die Temperatur beträgt deshalb durchschnittlich lediglich 17 °C."

Persönliche Daten

Name		Vorname	
Straße			
PLZ / Wohnort			
Geburtsdatum		Aktenzeichen	
Festnetz		Handy	

Antrag

☐ Ich bewohne folgende Wohnung
oder
☐ Wohnungswechsel: Ich werde zum u. g. Termin in folgende Wohnung einziehen

Straße	
PLZ / Wohnort	
Einzugsdatum in die Wohnung	

und beantrage die Übernahme von Renovierungskosten.

Abb. 2.1 Ausschnitt aus dem Antrag auf eine Beihilfe für Renovierungskosten. (Quelle: Jobcenter Kreis Warendorf, 2022)

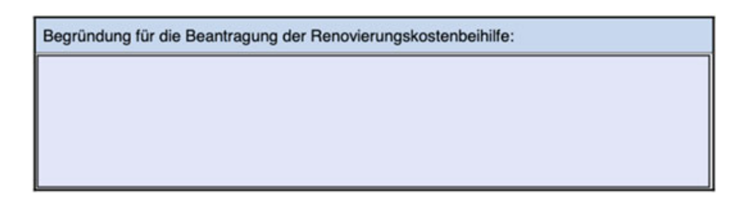


Abb. 2.2 Ausschnitt aus dem Antrag auf eine Beihilfe für Renovierungskosten. (Quelle: Jobcenter Kreis Warendorf, 2022)

► Versuchen Sie nun, die Frage in eine andere Form umzuwandeln, ohne dass Information möglicherweise verloren geht oder nicht abgefragt wird. Es sollte sich dabei um ein fixes Antwortschema statt einer offenen „Warum"-Frage handeln.

Die Übung zeigt, dass es nicht immer möglich ist, jeden Sachverhalt mit einem einfachen Antwortschema zu erfassen. Die Welt ist komplex und so sind es auch die zugehörigen Daten. Deshalb ist es wichtig zu verstehen, was bei der Erhebung und Verarbeitung von Daten zu beachten ist und wie man Daten verwenden und interpretieren kann.

2.2 Daten verstehen

Daten spielen im Kontext von Künstlicher Intelligenz eine entscheidende Rolle: ohne Daten keine KI. Daten werden daher auch häufig als das „Gold des 21. Jahrhunderts" bezeichnet (Focus, 2020), wobei hierzu direkt ergänzt werden sollte, dass Daten an sich keinen Mehrwert generieren, sondern erst wertvoll werden, wenn sie in Lösungen integriert sind (Stöger, 2017, S. 59). Während im Jahr 2008 die größten Unternehmen nach Marktkapitalisierung noch auf Öl und seine Derivate konzentriert waren – Exxon Mobile, Petro China, General Electrics – wird dieses Ranking mittlerweile von Unternehmen angeführt, deren Portfolios datenbasiert sind: Apple, Alphabet (Google) und Microsoft (Bünte, 2018, S. 2 f.). Auch im öffentlichen Sektor spricht man inzwischen von Data-Driven Government, in dessen Kontext die Potenziale datenbasierter Verwaltung ausgeschöpft werden sollen (Fadavian et al., 2019). Die Bundesregierung bezeichnet Daten sogar als die „Grundlage eines modernen Staates und einer mündigen Gesellschaft" (Bundesregierung, 2021a). Die Corona-Pandemie hat einmal mehr verdeutlicht, dass

zuverlässige Daten Grundlage für politische Maßnahmen sind, weshalb die Bundesregierung beschlossen hat, dass alle Ministerien sowie das Kanzleramt eigene Datenlabore unter der Leitung von Chief Data Scientists aufbauen sollen (Bundesregierung, 2021b). Was aber verbirgt sich überhaupt hinter dem Terminus Daten und wie kann man dies zu anderen Begrifflichkeiten abgrenzen?

Daten – Informationen – Wissen

Daten sind (alpha-)numerische Zeichenfolgen mit einer zugehörigen Syntax, welche durch Beobachtungen, Messungen, statistische Methoden o. ä. ermittelt wurden. Synonyme sind etwa Fakten oder Maße. Zentral ist die in der Regel digitale Verarbeitungsmöglichkeit von Daten (Abts & Mülder, 2017, S. 11). Daten unterliegen keinem Verschleiß, sie sind beliebig reproduzierbar und sie können über unterschiedliche Quellen verfügbar gemacht werden.

Informationen hingegen beziehen sich auf den Sinngehalt, der durch menschliche Interpretation von Daten entsteht (Deutscher Bundestag, 2020, S. 54). Informationen werden gewonnen durch die Auswertung von Daten. Hieraus entsteht Wissen. Letzteres kann als eine vom Menschen klassifizierte und interpretierte Auswertung von Informationen verstanden werden (Abts & Mülder, 2017, S. 12). Wissen ist die Grundlage für Entscheidungen und Handlungen. Heutzutage ist es weniger eine Herausforderung Informationen zu gewinnen, denn diese sind im Überfluss vorhanden (Stichwort *Information Overflow*), vielmehr geht es darum, Informationen in Wissen zu transformieren (vgl. von Rimscha, 2014, S. 27).

Struktur von Daten

Von der Struktur der Daten hängen weitere Verarbeitungsmöglichkeiten ab. Daher ist es von Bedeutung zu verstehen, inwiefern Daten unterschiedlich strukturiert sein können und was das für die Verarbeitung bedeutet. Man unterscheidet grundsätzlich strukturierte, semi-strukturierte und unstrukturierte Daten. Strukturierte Daten haben eine bestimmte Länge und ein vorgegebenes Format – sie besitzen eine bestimmte Struktur – beispielsweise die Kundennummer oder das Datum auf einer Online-Rechnung. Man spricht in diesem Kontext auch von traditionellen Daten, die in relationalen Datenbanken in Tabellenform gespeichert werden und einem vorgegebenen Datenmodell zugrunde liegen. Unstrukturierte Daten verfügen nicht über eine einheitliche Struktur, traditionelle Methoden der Datenanalyse können hierfür nicht angewendet werden. Zu dieser Kategorie gehören beispielsweise Berichte und Präsentationen, Fotos, Videos oder Kommentare. Es ist beispielsweise nicht möglich, aus Fotos einen Durchschnittswert zu errechnen, mit dem sinnvoll weitergearbeitet werden kann. Semi-strukturierte Daten sind zwischen diesen beiden Extremen zu verorten. Sie folgen keiner allgemein gültigen Struktur, enthalten aber

bestimmte Strukturinformationen, wie etwa die Kennzeichnung von Nachrichten in natürlicher Sprache mittels eines Hashtags. Etwa 20 % der weltweit vorhandenen digitalen Daten sind strukturiert, bei den restlichen 80 % handelt es sich um semi- oder unstrukturierte Daten (Heuberger-Götsch, 2016, S. 87).

Kategorisierung von Daten

Neben dem Grad der Strukturierung gibt es noch weitere Merkmale, anhand derer man Daten kategorisieren kann. Hierzu gehört etwa die Syntax, also die Regeln, nach denen eine formale Sprache mit einem vorgegebenen Zeichenvorrat gebildet wird. Man unterscheidet hierbei drei Arten von Syntax:

- numerisch: 0123456789
- alphabetisch: ABCDEFGHIJK
- alphanumerisch: AB5Z14TXM3

Darüber hinaus kann man Daten in der Art und Weise unterscheiden, in der sie nach außen in Erscheinung treten. Es kann sich beispielsweise um Texte, Zahlen, Bilder oder Audioaufnahmen handeln. Der zeitliche Bezug von Daten kann zwei Ausprägungen haben: entweder beziehen sich die Daten auf einen bestimmten Zeitpunkt (z. B. Anzahl der Corona-Neuinfektionen am 01.05.2020) oder auf einen Zeitraum (z. B. Anzahl der Corona-Neuinfektionen von März 2020 bis März 2022). Natürlich ist auch der statistische Aussagegehalt von Daten zu beachten. Man unterscheidet hierbei verschiedene Skalenniveaus (vgl. Fahrmeir et al., 2016, S. 16):

- **Nominal:** Daten können in keine logische oder natürliche Reihenfolge gebracht werden, z. B. Verkehrsmittel (Bus, Auto, Zug, E-Roller, …).
- **Ordinal:** Eine Rangfolge im Sinne von „ist größer“ oder „ist kleiner als“ ist möglich, der Abstand zwischen den Merkmalsausprägungen kann aber nicht interpretiert werden, z. B. Pflegestufen (Pflegestufe 1, Pflegestufe 2, Pflegestufe 3, …).
- **Kardinal:** Auch metrische Skala genannt, zusätzlich zu den bisherigen Eigenschaften sind die Differenzen zwischen den Merkmalsausprägungen hier interpretierbar, z. B. Gebühren in Euro oder die Anzahl denkmalgeschützter Gebäude.

Damit ein Computer die Daten „versteht“, muss festgelegt werden, um welchen Datentyp es sich handelt. Dies ist für die weitere Operationalisierung der Daten von Bedeutung. Es gibt eine lange Liste an Datentypen, diese variieren zudem zwischen den einzelnen Programmiersprachen (vgl. Sanella et al., 2022, S. 8). In

der folgenden Liste wird eine Auswahl an Datentypen vorgestellt, in Klammern ist die Bezeichnung des Datentyps, wie der Computer sie versteht):

- **Zeichen (CHAR):** Personalausweis
- **Ganzzahl (INTEGER):** 518024
- **Gleitkommazahl (FLOAT, DOUBLE):** $23{,}67 * 10^3$
- **Dezimalzahl (DECIMAL):** 95,14
- **Datum/Zeit (DATE, TIMESTAMP):** 15/05/2021
- **Boolean (BOOL):** TRUE/FALSE

Es ist nicht immer einfach, Daten genau zu kategorisieren und manchmal ist es auch möglich, dieselbe Variable durch unterschiedliche Datentypen darzustellen. Folgende Beispiele zeigen alle das Alter einer Person:

- 15.05.1982
- 15. Mai 1982
- 39 Jahre
- 15/05/1982
- 1982/05/15
- Neununddreißig Jahre
- 15.5.82

Das Wort „Neununddreißig" ist ein String bzw. eine Zeichenkette (CHAR), es wird von einem Computer ganz anders behandelt und kann in dieser Form nicht genauso verarbeitet werden wie etwa „15.05.1982" – also eine Datumsangabe (DATE). Manchmal ist auch nicht das genaue Geburtsdatum notwendig, sondern lediglich das Alter einer Person, etwa bei statistischen Erhebungen. Dann kann die Angabe 39 ausreichend sein, also ein INTEGER. Wichtig ist, sich im Vorfeld zu fragen, wie die Daten weiterverarbeitet werden sollen. Von großer Bedeutung ist außerdem, dass man sich für ein Format entscheiden muss und dieses dann auch durchgängig nutzt.

2.3 Übung zum Verständnis von Daten

Sehen Sie sich nun noch einmal das Formular zum Antrag auf eine Beihilfe für Renovierungskosten an (Abb. 2.3). Versuchen Sie, die dort im ersten Teil erhobenen personenbezogenen Daten möglichst präzise zu kategorisieren. Vielleicht fällt Ihnen auch bereits auf, an welcher Stelle dieses Formular zur weiteren Datenverarbeitung verbessert werden könnte.

Der „Name" ist beispielsweise alphabetisch, die Ausprägung der Daten tritt nach außen als Text auf und ist statistisch betrachtet dem nominalen Skalenniveau zuzuordnen. Es bietet sich an, die Angaben unter „Name" als Dateityp CHAR abzuspeichern. Durchlaufen Sie nun in dieser Art und Weise die anderen Angaben.

Beispiel

Im Feld PLZ/Wohnort wird es schwierig, denn hier tauchen zwei unterschiedliche Datentypen in einem Feld auf. Die Postleitzahl ist beispielsweise numerisch, der Wohnort hingegen alphabetisch. Die Postleitzahl tritt als Zahl in Erscheinung, der Wohnort als Text. Während die Postleitzahl als INTEGER gespeichert werden kann, wäre der Wohnort als CHAR zu speichern. Werden zwei unterschiedliche Datentypen in einem Feld gespeichert, kann dies im weiteren Verlauf zu Problemen führen. Als Lösung bietet es sich hier an, die beiden Angaben in getrennten Feldern abzufragen. Zusätzlich ist zu berücksichtigen, dass auch nicht-numerische Postleitzahlen auftreten können. Dies ist relevant wenn Personen aus anderen Staaten zuziehen oder in andere Staaten wegziehen. Darüber hinaus können bei Postleitzahlen führende

Persönliche Daten

Name		Vorname	
Straße			
PLZ / Wohnort			
Geburtsdatum		Aktenzeichen	
Festnetz		Handy	

Abb. 2.3 Ausschnitt aus dem Antrag auf eine Beihilfe für Renovierungskosten. (Quelle: Jobcenter Kreis Warendorf, 2022)

Nullen auftreten, die unter Umständen fehlerhaft verarbeitet werden. Zum Beispiel werden führende Nullen in Software zur Tabellenkalkulation ohne entsprechende Einstellungen automatisch entfernt.◀

2.4 Datenqualität

Unter Datenqualität wird ein Konzept verstanden, welches als Grundlage genutzt wird, um die Qualität von Daten objektiv zu bewerten. Seit spätestens Mitte der 1990er Jahre wird das Thema der Datenqualität systematisch wissenschaftlich untersucht. Dabei kommt es teilweise dazu, dass durch unterschiedliche Perspektiven auf Datenprobleme divergierende Definitionen derselben Dimension entstehen. Im Management bewertet man eine Dimension anders als mit einer Daten-orientierten Perspektive, mit einer reinen Anwendersicht beurteilt man ebenfalls anders. Dennoch existieren bestimmte Dimensionen, bei denen man sich einig ist, dass diese für die meisten Anwendungsbereiche von Bedeutung sind. Hierzu gehören (vgl. Hildebrand et al., 2018, S. 62–61):

- **Vollständigkeit:** Hiermit ist vor allem die Abwesenheit von NULL-Werten bzw. das Verhältnis von Non-NULL-Werten zur Gesamtheit aller Werte gemeint. NULL kann dabei für einen fehlenden Wert stehen, es kann sich aber auch um eine nicht zutreffende Angabe handeln, etwa wenn ein Bürger keine Angabe zum Geburtsnamen macht, weil dieser mit dem Familiennamen übereinstimmt. Vollständigkeit kann aber auch auf das Verhältnis zwischen den Daten in der Datenbank und denen in der Realwelt abzielen, beispielsweise ob alle im Straßenverkehr teilnehmenden Kraftfahrzeuge auch korrekt bei der entsprechenden Zulassungsstelle gemeldet sind. Es ist jedoch eine große Herausforderung, diese Form der Vollständigkeit zu überprüfen, da entweder zusätzliche Metadaten notwendig sind oder ein manueller Abgleich mit (zumindest einer Stichprobe) der Realwelt durchgeführt werden müsste.
- **Genauigkeit:** Dieser Aspekt beschreibt das Ausmaß, in dem Daten korrekt, zuverlässig und nachweislich fehlerfrei sind. Fraglich ist auch hierbei, inwiefern die Daten mit der Realwelt übereinstimmen – allerdings dieses Mal aus inhaltlicher Sicht. Wenn in einer Tabelle die Bearbeitungsgebühr für einen bestimmten Antrag mit 35 EUR aufgelistet wird, dann sollte dies auch die Summe sein, die in der Behörde tatsächlich anfällt und die formal vorgegeben ist. Unter Genauigkeit wird demnach die Differenz zwischen digitalen Daten und der realen Entsprechung verstanden, wobei diese möglichst gering sein

sollte. Eine Herausforderung ist dies teilweise bei semantischen Daten, die als Datentyp CHAR gespeichert werden und in der Regel Wörter aus natürlicher Sprache beinhalten. „München“ und „Munich“ sind von der Syntax her unterschiedlich, beides beschreibt aber dieselbe Stadt.
- **Konsistenz:** Unter Konsistenz versteht man das Ausmaß, in dem die Daten eines Systems den vorgegebenen Beschränkungen und Geschäftsregeln entsprechen. Hierbei kann es sich um klassische Datenbank-Vorgaben handeln, etwa dass die Kundennummer einmalig (unique) sein muss oder um Regeln wie etwa „Alter = heutiges Datum – Geburtsdatum“.
- **Aktualität:** Veraltete Daten führen zu Fehlern und reduzieren somit die Datenqualität. Es sollte daher sichergestellt werden, dass die verwendeten Daten aktuell sind. Es gilt dabei jedoch nicht grundsätzlich, dass ältere Daten, also mit einer größeren Zeitspanne zwischen der Erstellung und heute, automatisch weniger Wert hätten. Eine Kundennummer beispielsweise bleibt in der Regel gleich, unabhängig davon, wann sie erstellt wurde. Beträge für Sozialleistungen werden hingegen regelmäßig angepasst, etwa anhand der Inflationsrate. Es gilt, je häufiger Daten aktualisiert werden müssen, desto schneller altern sie. Gleichzeitig sind Daten, die nie modifiziert werden müssen (wie etwa die Kundennummer) immer aktuell. Mangelnde Aktualität wirkt sich also auch auf die Genauigkeit aus, was deutlich macht, dass die Dimensionen der Datenqualität häufig miteinander verwoben sind.

Daneben findet man in der Praxis noch weitere Kriterien, anhand derer die Datenqualität beurteilt wird, wie etwa die Integrität oder die Abstammung (Conformed Dimensions of Data Quality, 2021). Darüber hinaus sollten die Daten widerspruchsfrei sowie valide sein und Redundanz sollte vermieden werden. Das Fraunhofer-Institut empfiehlt, für die Gestaltung vertrauenswürdiger Künstlicher Intelligenz qualitative Anforderungen an die verwendeten Daten festzulegen. Dabei sollten mindestens folgende Aspekte berücksichtigt werden:

- technische Kriterien, wie etwa das Format oder die Größe der Datei;
- die Daten sollten vollständig sein;
- die Daten sollten mit der realen Welt übereinstimmen;
- Annotationen und Labels sollten korrekt sein;
- die Daten sollten relevant für den jeweiligen Anwendungsbereich sein;

- ein Zugriff auf die Daten sowie die zugehörigen Metadaten sollte sichergestellt bzw. die Daten sollten jederzeit verfügbar sein (Fraunhofer, 2021, S. 93).

Wichtig ist jedoch, dass bei der Qualitätsbewertung nicht nur technische Aspekte, sondern auch der Inhalt und die Verständlichkeit der Daten berücksichtigt werden. Möchte man etwa ein KI-System trainieren, Bilder von Hunden zu erkennen, dann benötigt man einen Datensatz mit entsprechenden Bildern. Hat der Datensatz aber nur Bilder von ein und demselben Hund, dann ist die Qualität gering – selbst wenn alle technischen Vorgaben erfüllt sind. Das KI-System kann anhand solcher Daten nicht lernen, allgemein Hunde auf Fotos zu erkennen. Es wird lediglich diesen einen Hund identifizieren können.

2.5 Beziehungen zwischen Daten

Man differenziert bei Daten zwischen unabhängigen und abhängigen Variablen. Abhängige Variablen hängen hierbei von unabhängigen Variablen ab und werden von diesen beeinflusst. Bei der sogenannten Regressionsanalyse wird beispielsweise die Beziehung dieser Variablen zueinander untersucht nach dem Schema „je mehr x, desto mehr y“ oder „je mehr x, desto weniger y“. Hier hängt der Wert von y davon ab, wie der Wert von x ist. Damit ist y abhängig von x und x ist unabhängig.

Zum einen sollen die abhängigen Parameter auf Basis der unabhängigen prognostiziert und zum anderen der Grad einer Korrelation festgestellt werden. Das Prinzip der Regressionsanalyse wird im dritten Kapitel dieses Moduls genauer erläutert. Wenn etwa die Dauer einer Altbausanierung die Variable ist, die erklärt werden soll, dann sind die Wetterbedingungen während der Bauzeit eine unabhängige Variable, die auf das Ergebnis Einfluss nimmt. Diese Beziehung erscheint zwar auf dem ersten Blick einfach und nachvollziehbar, allerdings sollte man eine berühmt gewordene Weisheit aus der Statistik immer im Hinterkopf behalten:

▶ Korrelation ist ungleich Kausalität!

Es ist möglich, dass zwischen Variablen ein statistischer Zusammenhang (z. B. eine Korrelation) besteht, dies bedeutet aber nicht automatisch, dass eine Ursache-Wirkung Beziehung (Kausalität) vorliegt (vgl. Fahrmeir et al., 2016, S. 140 f.). Ein in diesem Kontext häufig aufgeführtes Beispiel ist das der Störche und Babys. Demnach würde die Anzahl der Störche mit der Anzahl der Geburten korrelieren, also je mehr Störche desto mehr Babys. Man könnte nun also voreilig

annehmen, die Anzahl der Störche sei die Ursache für die Anzahl der Babys. Allerdings war nicht nur die Erhebung von lediglich zwei Variablen für eine Untersuchung äußerst schwach, es wurde vielmehr eine dritte Variable schlicht ignoriert: die geographische Lage. Denn tatsächlich war der ländliche Raum Ursache für sowohl die Anzahl der Störche als auch für die Anzahl der Babys. Die beiden Variablen selbst – also die Störche und die Babys – standen jedoch nicht in kausaler Beziehung zueinander, sondern korrelierten aufgrund einer dritten Variable: dem ländlichen Raum.

Schließlich sei noch auf die Bezeichnungen hingewiesen. Wie so häufig in der IT-Welt gibt es nicht nur zwei eindeutige Begrifflichkeiten für diese beiden Arten von Variablen. Unabhängige Variablen werden auch als Input-Feature, Regressor, erklärende oder exogene Variable, Prädiktor oder Faktor bezeichnet. Eine abhängige Variable wird auch als Regressand, endogene oder erklärte Variable, Ziel- oder Prognosevariable oder als zu erklärende Variable benannt.

2.6 Die Verzerrung von Daten (Bias)

KI-Systeme werden in der Regel mit historischen Daten trainiert. Das bedeutet, dass diese Daten alle in der Vergangenheit entstanden sind. Diese Daten sind nicht automatisch neutral, wert- oder vorurteilsfrei – auch dann nicht, wenn die Qualität der Daten als hoch eingestuft wird. Es kann beispielsweise sein, dass bestimmte Gruppen in den Daten nicht oder kaum vorkommen. Eine solche Datenverzerrung, häufig als *Bias* bezeichnet, beeinflusst das Ergebnis des KI-Systems signifikant und es können Benachteiligungen entstehen. In den USA wurde beispielsweise ein KI-System eingesetzt, das Patientinnen und Patienten mit besonderem Pflegebedarf identifizieren sollte. In einer Studie über dieses System wurde jedoch festgestellt, dass bei afroamerikanischen Menschen seltener ein zusätzlicher Pflegebedarf identifiziert wurde als bei Weißen – und das bei gleicher Krankheitsschwere. Eine manuelle Überprüfung ergab, dass die doppelte Anzahl an afroamerikanischen Menschen den Pflegebedarf benötigt hätte. Wie kam es nun, dass das KI-System diskriminierte, obwohl weder Hautfarbe noch ethnische Zugehörigkeit Daten waren, mit denen es trainiert wurde?

Der Grund hierfür ist, dass der Algorithmus die zu erwartenden Kosten für das Gesundheitssystem berücksichtigte statt der tatsächlich vorliegenden Krankheit und dem zugehörigen Pflegebedarf. Aufgrund diverser Faktoren gaben und geben die USA weniger Geld aus für die Behandlung von schwarzen als für die

von weißen Menschen. Diese Prognose geringerer Kosten wurde mit der Notwendigkeit für Pflegebedarf vom System gleichgesetzt, was zu den angesprochenen verzerrten Ergebnissen führte (vgl. Obermeyer et al., 2019).

Ein solcher Bias kann unterschiedliche Ursachen haben, die Daten können im Vorfeld nicht vollständig oder durch Manipulation verzerrt sein, es kann aber auch sein, dass die Daten mit der Realität übereinstimmen, diese jedoch Ungerechtigkeiten und Benachteiligung enthält. Es ist ein spannendes und wichtiges Thema, welches bei jeder Konzeption eines KI-Systems mit bedacht werden sollte. In Kap. 11 „KI und Ethik“ wird dieses Thema detailliert beleuchtet.

2.7 Aufgaben zum eigenen Anwendungsfall

Sie haben nun gelernt, dass der Begriff Daten vielschichtig ist. Daten sind die Grundlage für jedes KI-System. Überlegen Sie nun, welche Daten Sie für Ihr Projekt benötigen:

- Was ist Ihre Datenquelle bzw. welche Daten möchten Sie nutzen?
- Handelt es sich um Daten, die intern in Ihrer Organisation vorliegen oder benötigen Sie (entweder zusätzlich oder ausschließlich) Daten externer Quellen?
- Handelt es sich um strukturierte oder unstrukturierte Daten?
- Wählen Sie aus Ihrem Datensatz zwei bis drei Variablen und beschreiben Sie deren Syntax, Erscheinung, zeitlichen Bezug, das Skalenniveau und den Datentyp.

Literatur

Abts, D., & Mülder, W. (2017). *Grundkurs Wirtschaftsinformatik: Eine kompakte und praxisorientierte Einführung* (9., erweiterte und aktualisierte Aufl.). Springer Vieweg. https://doi.org/10.1007/978-3-658-16379-2.

Bundesregierung. (2021a). *Die Bundesregierung gründet Datenlabore und integriert Chief Data Scientists in alle Bundesministerien.* Webseite der Bundesregierung | Startseite. https://www.bundesregierung.de/breg-de/aktuelles/die-bundesregierung-gruendet-datenlabore-und-integriert-chief-data-scientists-in-alle-bundesministerien-1944226. Zugegriffen: 15. Okt. 2022.

Bundesregierung. (2021b). *Open-Data-Strategie der Bundesregierung*. Webseite der Bundesregierung | Startseite. https://www.bundesregierung.de/breg-de/aktuelles/open-data-strategie-1939808. Zugegriffen: 15. Okt. 2022.

Bünte, C. (2018). Künstliche Intelligenz – Die Zukunft des Marketing: Ein praktischer Leitfaden für Marketing-Manager. *Springer Gabler*. https://doi.org/10.1007/978-3-658-23319-8.

Conformed Dimensions of Data Quality. (o. J.). German list of underlying concepts | Conformed dimensions of data quality. http://dimensionsofdataquality.com/content/german-list-underlying-concepts. Zugegriffen: 25. Apr. 2022.

Deutscher Bundestag. (2020). Deutscher Bundestag–Enquete-Kommission „Künstliche Intelligenz" – Gesellschaftliche Verantwortung und wirtschaftliche, soziale und ökologische Potenziale. Drucksache 19/23700. https://www.bundestag.de/webarchiv/Ausschuesse/ausschuesse19/weitere_gremien/enquete_ki. Zugegriffen: 15. Okt. 2022.

Dudenredaktion. (o. J.). Daten. In *Duden*. https://www.duden.de/rechtschreibung/Daten. Zugegriffen: 15. Apr. 2022.

Fadavian, B., Franzen-Paustenbach, D., Rehfeld, D., Schmitt, M., Schweikart, D., & Djeffal, C. (2019). Data Driven Government. *Berichte Des NEGZ*, 1. https://doi.org/10.30418/2626-6032.2019.04.

Fahrmeir, L., Heumann, C., Künstler, R., Pigeot, I., & Tutz, G. (2016). *Statistik: Der Weg zur Datenanalyse*. Springer-Verlag.

Focus Online. (2020). Gold des 21. Jahrhunderts: Was Anleger über die Megatrends KI und Big Data wissen müssen. https://www.focus.de/finanzen/boerse/geldanlage/das-gold-des-21-jahrhunderts-big-data-und-ki-was-anleger-nicht-ueber-diesen-megatrend-wissen_id_12418970.html. Zugegriffen: 15. Okt. 2022.

Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS (Hrsg.). (2021). Leitfaden zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz. https://www.iais.fraunhofer.de/content/dam/iais/fb/Kuenstliche_intelligenz/ki-pruefkatalog/202107_KI-Pruefkatalog.pdf. Zugegriffen: 15. Okt. 2022.

Heuberger-Götsch, O. (2016). Der Wert von Daten aus juristischer Sicht am Beispiel des Profiling. In D. Fasel & A. Meier (Hrsg.), *Big Data: Grundlagen, Systeme und Nutzungspotenziale* (S. 83–105). Springer Fachmedien. https://doi.org/10.1007/978-3-658-11589-0_5.

Hildebrand, K., Gebauer, M., Hinrichs, H., & Mielke, M. (2018). *Daten-und Informationsqualität*. Springer Fachmedien Wiesbaden.

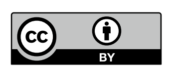
Jobcenter Kreis Warendorf. (2022). Antrag auf eine Beihilfe für Renovierungskosten gemäß § 22 Absatz 1 SGB II. https://www.jobcenter-warendorf.de/fileadmin/jobcenter/Antragsunterlagen/Sonstige_Antr%C3%A4ge/10_Antrag_auf_Renovierungskosten.pdf. Zugegriffen: 15. Okt. 2022.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453. https://doi.org/10.1126/science.aax2342.

Sannella, D., Fourman, M., Peng, H., & Wadler, P. (2022). *Introduction to computation: Haskell, logic and automata.* Springer Nature.

Stöger, R. (2017). Umsetzung der Digitalisierung. Fazit 1.0 in der Neuen Welt. *Zeitschrift für Organisationsentwicklung*, *36*(1), 58–64.

Von Rimscha, M. (2014). *Algorithmen kompakt und verständlich.* Springer Fachmedien Wiesbaden.



Check for updates

3 Grundlagen: Verarbeitung

Zusammenfassung

Im vorigen Kapitel wurde das Thema Daten aus verschiedenen Perspektiven behandelt und erörtert. Nun geht es um die Frage, wie ein KI-System Daten verarbeitet. Hierfür werden zunächst die unterschiedlichen Trainingsmethoden vorgestellt: überwachtes Lernen (3.2), unüberwachtes Lernen (3.3) sowie bestärkendes Lernen (3.4). Es folgt die Einführung künstlicher neuronaler Netze (3.5), deren Entwicklung momentan rasant voranschreitet. Die beiden folgenden Abschnitte befassen sich mit den Verfahren Support Vector Machine (3.6) sowie lineare und logistische Regression (3.7). Das angeeignete Wissen kann in einer Übung reflektiert (3.8) und im eigenen Anwendungsfall eingesetzt werden (3.9).

3.1 Einleitung

Das vorige Kapitel befasste sich mit dem Thema Daten als zentrale Grundlage für ein KI-System. Aber was passiert nun, wenn man Daten einem KI-System zur Verfügung stellt bzw. ein KI-System mit diesen Daten trainiert? Wie verarbeitet das System die Daten? Ein KI-System lernt nicht auswendig, es lernt grundsätzliche Strukturen, weshalb das System nicht nur mit den bereits bekannten Trainingsdaten Ergebnisse liefert, sondern auch mit zuvor völlig unbekannten, neuen Daten einen Output generiert. Im Folgenden werden drei maßgebliche Verfahren des maschinellen Lernens vorgestellt, das überwachte, unüberwachte und bestärkende Lernen (vgl. Seegerer et al., 2020 sowie Russel & Norvig, 2012, S. 811).

M. Heine et al., *Künstliche Intelligenz in öffentlichen Verwaltungen*, Edition eGov-Campus, https://doi.org/10.1007/978-3-658-40101-6_3

3.2 Überwachtes Lernen (Supervised Learning)

Ausgangslage für das überwachte Lernen sind Daten, die ein Label enthalten. Als Label bezeichnet man die Information oder das Ergebnis, mit dem der Input verbunden werden soll. Das könnten am Beispiel der öffentlichen Verwaltung auch „Angenommen" oder „Abgelehnt" bei Bescheiden sein. Ein anderes Beispiel sind viele handschriftlich erstellte Buchstaben, die dann mit dem korrekten Label gekennzeichnet sind, damit die Maschine versteht, welcher Buchstabe hinter der Schrift steht. In Abb. 3.1 etwa sind verschiedene handschriftlich geschriebene Buchstaben mit dem jeweils für die Maschine lesbaren Buchstaben gekennzeichnet.

Die Daten werden anschließend in Trainings- und Testdaten aufgeteilt, wobei die Menge der Trainingsdaten deutlich größer ist. Das System lernt nun anhand der Trainingsdaten die Strukturen und charakteristischen Merkmale der einzelnen



Abb. 3.1 Buchstaben mit Label

Buchstaben. Anschließend wird anhand der Testdaten geprüft, wie präzise das System die Daten korrekt zuordnet.

Durch überwachtes Lernen können hauptsächlich zwei Aufgaben erfüllt werden: Klassifikation und Regression. Ersteres bedeutet beim Beispiel der Buchstaben, dass das System für jeden Buchstaben eine Kategorie anhand ähnlicher Merkmale erstellt. Ein „C“ hat beispielsweise keine gerade Linie und besteht nur aus einer durchgängigen, gebogenen Linie. Dementsprechend ordnet das System die Zeichen, welche diese Merkmale bestätigen, der Kategorie „C“ zu. Ebenso könnte das System beispielsweise prüfen, ob ein bestimmtes Antragsformular vollständig ausgefüllt ist oder nicht.

Bei der Regression wird hingegen ein kontinuierlicher numerischer Wert ermittelt, etwa wie hoch das Einkommen ist – z. B. in Abhängigkeit vom Alter. Hierbei muss allerdings ein besonderes Phänomen verhindert werden, das sogenannte *Overfitting* (zu Deutsch in etwa Überanpassung). *Overfitting* bedeutet, das System lernt die Trainingsdaten auswendig bzw. es erkennt keine allgemeinen Strukturen, sondern richtet die Auswertung eng an den Trainingsdaten aus (vgl. Russel & Norvig, 2012, S. 821 f.). Das kann zum Beispiel passieren, wenn nur Personen mit hohem Alter in den Trainingsdaten sind, in der wirklichen Anwendung aber auch andere Altersbereiche auftauchen. Das führt dazu, dass das System unbekannte Daten nicht angemessen verarbeitet und somit fehlerhafte Ergebnisse erzeugt werden. Gleichzeitig ist es aber auch möglich, dass ein Zustand des *Underfitting* (etwa Unteranpassung) entsteht und das System den Zusammenhang zwischen den Daten nicht genug lernen kann. Dazwischen gibt es einen Zustand, der idealerweise erreicht werden sollte. In diesem ist die Auswertung der Daten ausgeglichen und das System hat weder die Trainingsdaten auswendig gelernt noch zu wenig Kenntnisse über allgemeine Strukturen in den Daten. Abb. 3.2 veranschaulicht diese unterschiedlichen Formen der Anpassung.

Beim überwachten Lernen wird also zunächst ein Datensatz benötigt, dessen Daten über ein Label verfügen. Schon diese erste Voraussetzung ist nicht immer einfach zu erfüllen. Darüber hinaus benötigt man eine gewisse Menge an Daten, um *Underfitting* zu vermeiden, und eine gewisse Qualität und Diversität der Daten, um *Overfitting* zu verhindern. Außerdem ist es von großer Bedeutung, vorab das Ziel des Systems zu definieren: Sollen Daten klassifiziert werden oder wird eine Regression benötigt? Mithilfe der Testdaten wird überprüft, wie präzise das System mit unbekannten Daten umgeht.

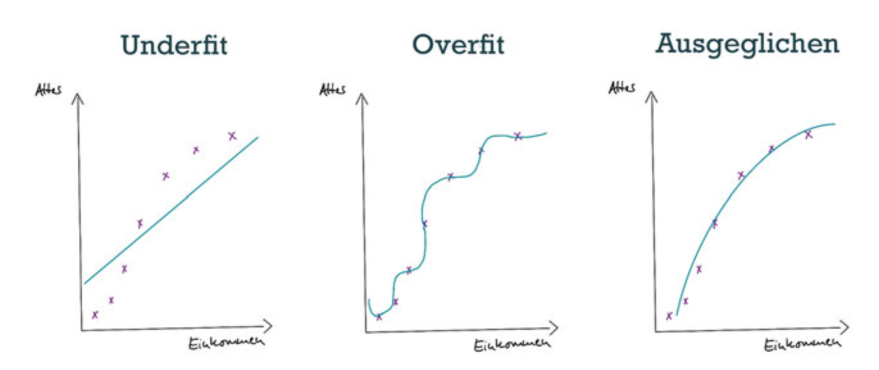


Abb. 3.2 Problematische Entwicklungen beim Trainieren des KI-Systems

3.3 Unüberwachtes Lernen (Unsupervised Learning)

Ausgangslage für das unüberwachte Lernen sind große Datenmengen, die kein Label enthalten, beispielsweise verschiedene Antragsformulare im pdf-Format oder Videos über Parlamentsdebatten. Dadurch kann das unüberwachte Lernen auch dort eingesetzt werden, wo es noch keine definierten Zielwerte gibt (vgl. Kreutzer & Sirrenberg, 2019, S. 7), etwa wenn einem Sachbereich viele verschiedene Aufgaben neu zugeordnet werden und überlegt werden muss, an welchen Stellen man diese gruppieren kann. Wie findet man nun also eine zusammenhängende Muster in diesen Daten? Wie kann man diese Daten effektiv und sinnvoll gruppieren? Solche Aufgaben können mithilfe von unüberwachtem Lernen erfüllt werden (vgl. Buxmann & Schmidt, 2019, S. 10).

Das Gruppieren von Daten wird auch als Clustering bezeichnet. Innerhalb der Cluster befinden sich Daten mit ähnlichen Ausprägungen. Die oben aufgeführten Antragsformulare könnten etwa in verschiedene Cluster eingeteilt werden, in denen zum Beispiel Anträge für Baumaßnahmen, Anträge für Wohngeld oder Anträge für Elterngeld gruppiert sind. Es ist möglich, die Anzahl der gewünschten Cluster vorab festzulegen, wobei zu beachten ist, dass eine hohe Anzahl an Clustern kleinere Gruppen mit mehr Granularität erzeugt, eine geringe Anzahl an Clustern hingegen zu größeren Gruppen mit geringerer Granularität führt.

Im vorigen Abschnitt wurde das Auslesen handschriftliche Buchstaben mittels überwachten Lernen trainiert. Dies war möglich, weil der Datensatz für jeden Buchstaben ein Label enthielt. Wie verhält es sich nun, wenn wir einen Datensatz handschriftlicher Buchstaben vorliegen haben, diese aber nicht direkt



Abb. 3.3 Buchstaben ohne Label

annotiert sind und somit von einer Maschine nicht zugeordnet werden können (wie in Abb. 3.3)?

Unüberwachtes Lernen ist hier eine Möglichkeit, der Maschine ein Verständnis für die unterschiedlichen Buchstaben anzutrainieren. Die Anzahl der Cluster wird mit der Anzahl der Buchstaben gleichgesetzt, die Maschine gruppiert nun die Buchstaben anhand ähnlicher Ausprägungen in dieselben Cluster (Abb. 3.4).

Im Gegensatz zum überwachten Lernen ist es beim unüberwachten Lernen deutlich schwieriger, die Zuverlässigkeit des Algorithmus zu überprüfen, weil es keinen Test-Datensatz gibt, durch den die Ergebnisse des Systems eindeutig bewertet werden können. Daher ist es in solchen Fällen besonders wichtig, ein Verständnis für die Daten zu entwickeln und die Ergebnisse des Systems kritisch zu hinterfragen.

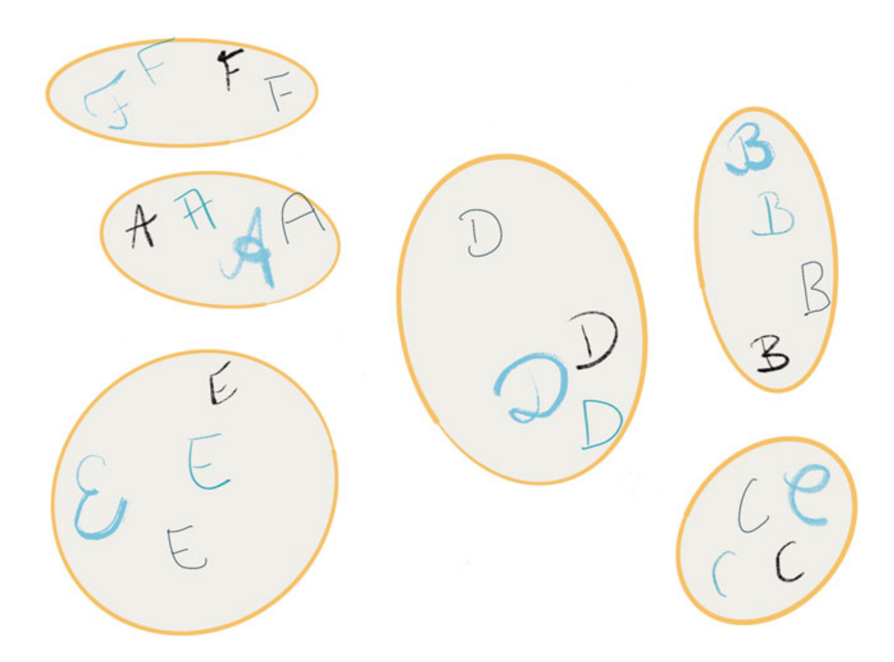


Abb. 3.4 Cluster ähnlicher Buchstaben

3.4 Bestärkendes Lernen (Reinforcement Learning)

Verstärkendes oder bestärkendes Lernen (englisch: Reinforcement Learning) unterscheidet sich in seiner Form von den vorherigen Verfahren. Das KI-System erlernt die korrekte Zuordnung von Daten „spielerisch“, indem es für korrekte Antworten belohnt und für inkorrekte Antworten bestraft wird. Dieses Verfahren gibt es auch in der Psychologie und wird dort als operantes Konditionieren bezeichnet. In der Regel liegen bei diesen Anwendungsfällen keine großen Datenmengen vor. Vielmehr lernt das System durch Erfahrung, es sammelt die Rückmeldungen der jeweiligen Outputs und versucht langfristig die Belohnung zu maximieren.

Ein klassisches Anwendungsfeld für bestärkendes Lernen ist die Robotik. Die Aufgaben von Robotern sind in der Regel komplex, sie lassen sich nicht durch Programmieren endgültig festlegen. Außerdem liegen für die Anwendungsfälle von Robotern häufig keine Trainingsdaten vor. Daher muss der Roboter durch

Versuchen erlernen, welche Situationen zum Erfolg führen und welche ein Irrtum sind (vgl. Ertel, 2016, S. 313).

3.5 Künstliche Neuronale Netze

Viele erfolgreiche Erfindungen imitieren Funktionsweisen aus der Natur. Für die Entwicklung eines intelligenten Systems hat man sich am menschlichen Gehirn – insbesondere am Aufbau der 10 bis 100 Mrd. Nervenzellen bzw. Neuronen – orientiert. Entstanden ist dabei das Konstrukt des künstlichen neuronalen Netzes (vgl. Russel & Norvig, 2012, S. 845–856).

Wie auch bei den Verfahren des maschinellen Lernens erhält das künstliche neuronale Netz (KNN) als Input Daten, welche es verarbeitet. Anschließend liefert das Netz als Output ein bestimmtes Ergebnis. Was beim Menschen durch organische Prozesse funktioniert, wird bei einem KNN durch mathematische Funktionen ausgeführt. Die Struktur eines neuronalen Netzes unterteilt sich dabei grundsätzlich in drei Schichten: der Input-Schicht, der „versteckten" Schicht (häufig als *Hidden Layer* bezeichnet) und der Output-Schicht (siehe Abb. 3.5). In jeder Schicht befinden sich Neuronen, welche mit einer unterschiedlichen Anzahl an Neuronen aus anderen Schichten verbunden sind. Zwischen diesen Neuronen werden Werte hin und her übermittelt. Diese Verbindungen entwickeln unterschiedliche Gewichtungen, in Abhängigkeit der Bedeutung des Austauschs. Jedes Neuron enthält eine sogenannte Aktivierungsfunktion, diese besteht aus der Summe der gewichteten Ausgabewerte der vorigen Neuronen, die mit ihm verbunden sind. Wenn Informationen nicht nur weitergeleitet, sondern auch zurückgeleitet bzw. rückgekoppelt werden können, spricht man von rekurrenten Netzen. Dies ist beispielsweise notwendig, wenn das KNN während des Lernens auf Informationen zurückgreifen soll.

Dieses neuronale Netz kann man nun überwacht oder unüberwacht trainieren, ebenso kann man Reinforcement Learning nutzen. Es gibt verschiedene „Architekturen" künstlicher neuronaler Netze, diese sich in der Funktionalität der einzelnen Neuronen sowie in der Gesamtfunktionalität unterscheiden. Im Bereich der Bilderkennung werden beispielsweise konvolutionale Netze eingesetzt (*Convolutional Neural Networks*, CNN), während man bei der Zeitreihenanalyse gern auf sogenannte LSTM (*Long Short-Term Memory*) zugreift. Welche Art von KNN passend ist, hängt also von dem jeweiligen Anwendungsfall und auch von den vorliegenden Daten ab. Wenn man wieder an das Beispiel des Auslesens handschriftlich erstellter Schrift denkt und hierfür genügend Daten vorliegen, kann man ein CNN mit diesen trainieren. Wenn eine Vielzahl an KNNs genutzt wird

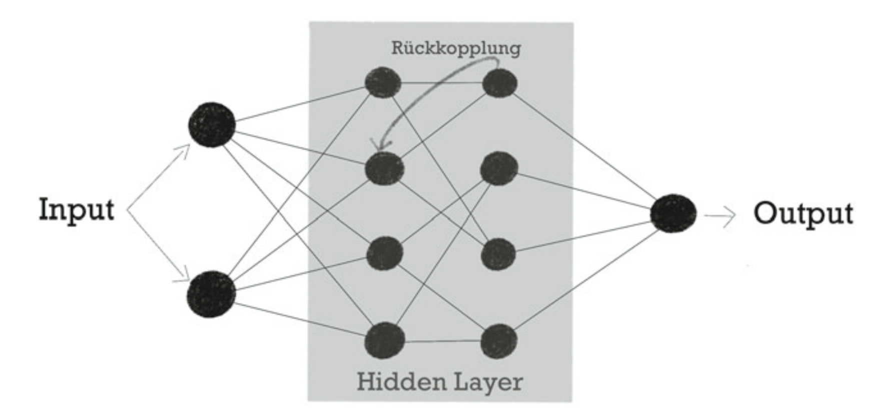


Abb. 3.5 Schematische Darstellung eines künstlichen neuronalen Netzes

und die Anzahl der Hidden Layer steigt, spricht man auch von tiefen neuronalen Netzen oder von *Deep Learning* (vgl. Ertel, 2016, S. 300).

Künstliche neuronale Netze sind komplexe Gebilde – um ein vertieftes Verständnis über die Funktionsweise zu erhalten, ist es notwendig, sich mit der zugrunde liegenden Mathematik und Statistik auseinanderzusetzen. Dieses Buch kann hier nur einen groben Überblick bieten. Angesichts der vielfältigen Projekte und Publikationen derzeit lohnt sich eine Vertiefung über dieses Buch hinaus.

3.6 Support Vector Machine

Eine Support Vector Machine (SVM) ist keine Maschine im physischen Sinne, vielmehr handelt es sich um einen Algorithmus, mit dessen Hilfe man Daten klassifizieren kann (vgl. Russel & Norvig, 2012, S. 863 f.). Eine SVM sucht hierfür eine „Hyperebene", mittels derer die Daten optimal voneinander getrennt werden. Man kann diese Hyperebenen als Entscheidungsgrenzen verstehen, anhand derer entschieden wird, in welche Klasse ein Datenpunkt fällt.

Wie in den Kennzeichnungen von Abb. 3.6 zu erkennen ist, gibt es häufig mehrere Möglichkeiten, diese Hyperebene zu platzieren. Ziel ist es, dass zwischen Hyperebene und dem nächsten Punkt auf beiden Seiten der maximal mögliche Abstand erreicht wird. Die rechts abgebildete Hyperebene ist demnach präziser als die auf der linken Seite. Je größer der Abstand zwischen den Datenpunkten

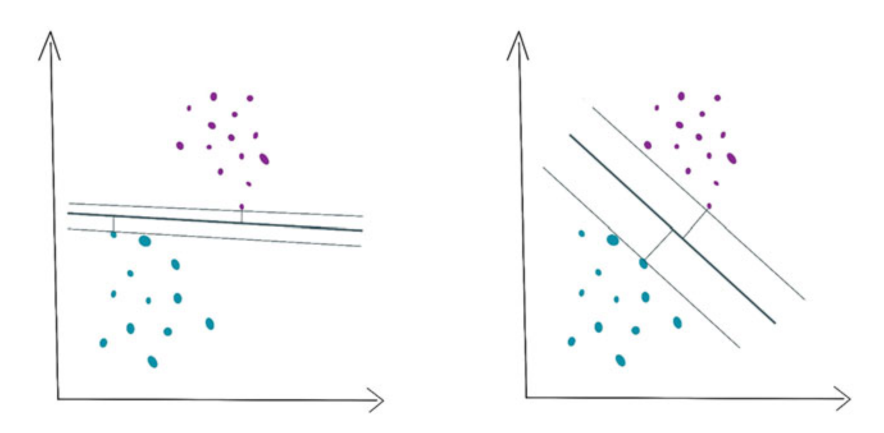


Abb. 3.6 Schematische Darstellung der Hyperebene einer SVM

und der Hyperebene, desto geringer ist die Wahrscheinlichkeit, dass Daten fehlerhaft klassifiziert werden. In der Abbildung werden die Daten in zwei Gruppen eingeteilt, es gibt also zwei unterschiedliche Arten von Input-Daten (hier türkis und violett). Die SVM kann aber auch eingesetzt werden, wenn es mehr als zwei Datenklassen gibt, dann gestaltet sich allerdings die Visualisierung schwieriger.

3.7 Lineare und logistische Regression

Bei der linearen Regression wird ein kontinuierlicher numerischer Wert ermittelt, etwa wie hoch das Einkommen in Abhängigkeit vom Alter ist (vgl. Russel & Norvig, 2012, S. 835–843). Die Korrelation zwischen einer abhängigen Variable und einer oder mehreren unabhängigen Variablen soll ermittelt werden. Für eine lineare Regression wird das Verfahren des überwachten Lernens angewendet, es werden dementsprechend Daten mit Label benötigt. Anschließend versucht das System, ein Muster (Pattern) in den Daten zu erkennen, bei der linearen Regression in Form einer „je mehr X, desto mehr Y“ oder „je mehr X, desto weniger Y“ Beziehung. Y ist dabei die abhängige Variable, also die unbekannte Variable, die von Interesse ist und bestimmt werden soll. X ist eine oder es sind mehrere Variablen, welche Einfluss auf Y haben und die bekannt sind.

Beispiel

Es soll geschätzt werden, wie viel Personal für Straßenreinigung in den Herbstmonaten benötigt wird, um das Laub von den Straßen zu entfernen. Die zugrunde liegenden Daten sind die Anzahl der Bäume je Straße und der Personalaufwand der vorherigen Jahre, veranschaulicht in Abb. 3.7.

Die Frage lautet nun, wie viele Arbeitsstunden für eine bestimmte Straße anfallen werden, um darauf aufbauend den Personalbedarf zu planen. Die Daten werden also zunächst in Trainings- und Testdaten aufgeteilt und das System erlernt die optimale Regressionsgerade zwischen den Datenpunkten. Um dies zu erreichen, muss der Abstand zwischen jedem realen Datenpunkt und dem zugehörigen vorhergesagten Punkt so klein wie möglich sein.

In Abb. 3.8 erkennt man nun die Regressionsgerade. Der Abstand zu den realen Datenpunkten wurde eingezeichnet, dieser muss in Summe so klein wie möglich sein. Wenn wir die Maschine nun fragen, wie viele Arbeitsstunden für die Reinigung einer Straße mit X Bäumen notwendig sind, wird die Antwort der zugehörige Y-Wert auf der Regressionsgeraden sein.

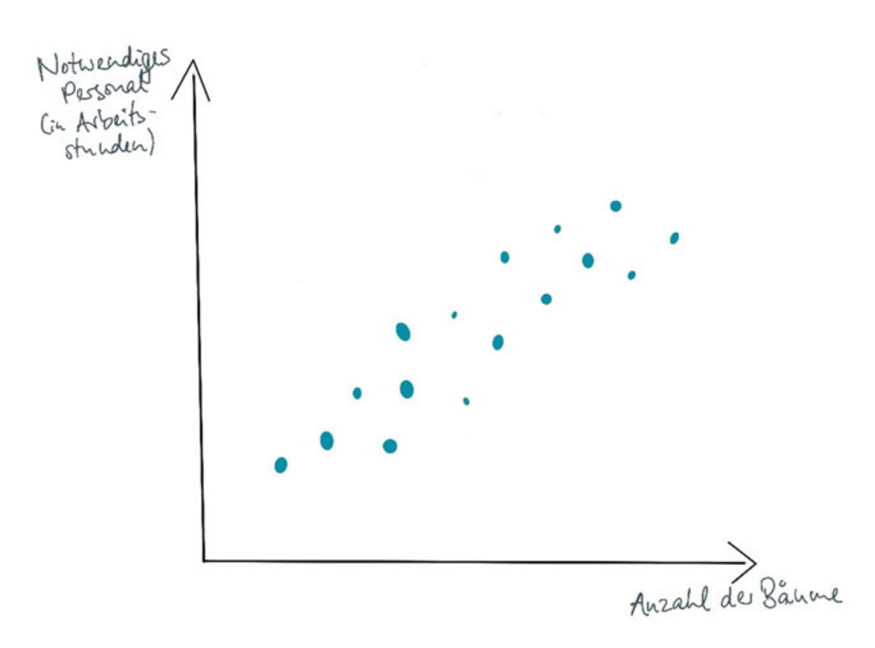


Abb. 3.7 Notwendiges Personal je Anzahl der Bäume

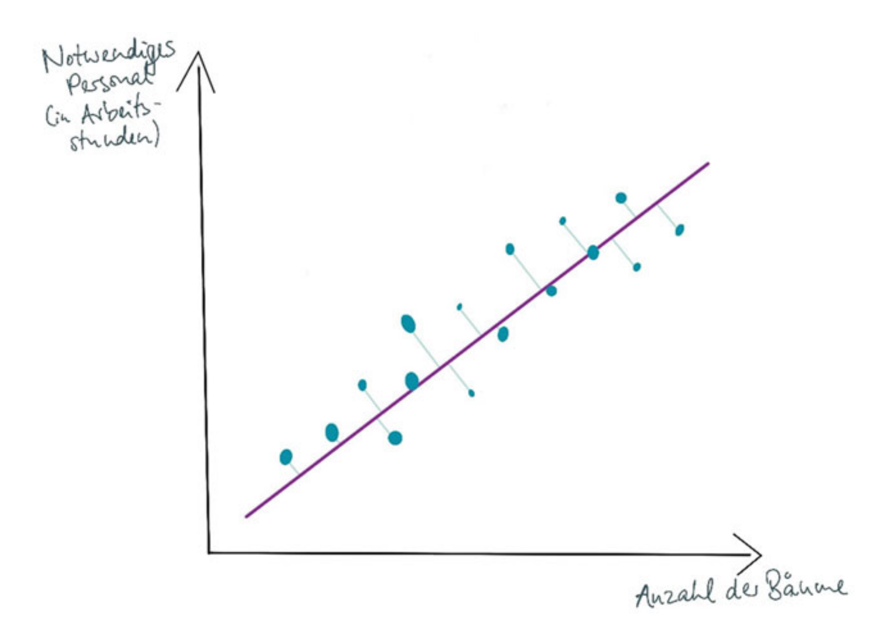


Abb. 3.8 Regressionsgerade zwischen benötigtem Personal und Anzahl der Bäume

◀

Bei der logistischen Regression geht es hingegen um Wahrscheinlichkeiten, konkret um die Wahrscheinlichkeit, ob ein Ereignis zutrifft oder nicht zutrifft (vgl. Russel & Norvig, 2012, S. 843–844). In der Regel wird hier eine Sigmoid-Funktion angewendet, die einen Wert zwischen 0 und 1 annehmen kann, s-förmig verläuft, einen Wendepunkt bei 0,5 hat und wie in Abb. 3.9 dargestellt aussieht.

Um einen geeigneten Anwendungsfall für logistische Regression zu generieren, wird also zunächst eine abhängige Variable, die dichotom ist, benötigt. Das bedeutet, es gibt zwei mögliche Kategorien (ja/nein, TRUE/FALSE usw.) – es sind also nominalskalierte Daten (zur Skalierung siehe Kap. 2). Es gibt zwar auch sogenannte multinomiale logistische Regressionen, bei denen die abhängige Variable mehr als zwei Ausprägungen haben kann, diese sollen an dieser Stelle jedoch vernachlässigt werden.

Im Gesundheitswesen könnte man eine logistische Regression beispielsweise einsetzen, um die Frage zu beantworten, ob eine Probe auf eine Krebserkrankung

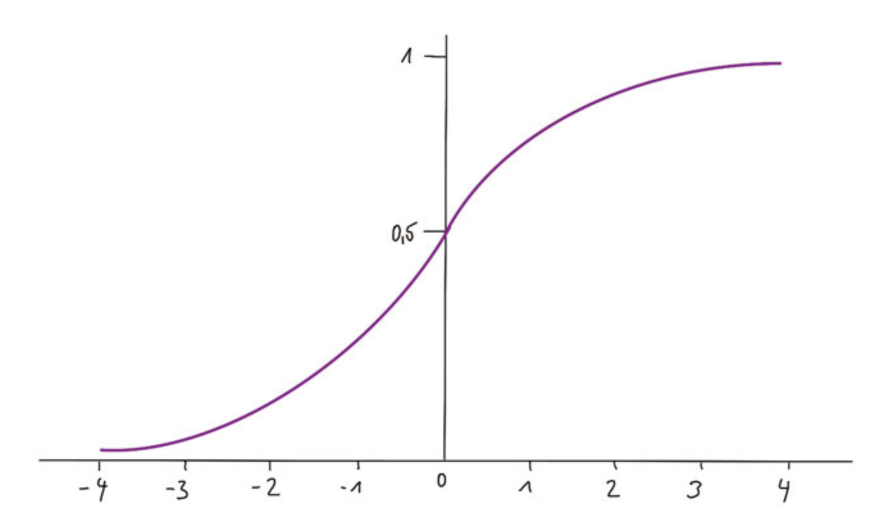


Abb. 3.9 Schematische Darstellung einer Sigmoid-Funktion

hinweist oder nicht. In einem E-Mail-Postfach könnte man mit der Wahrscheinlichkeit rechnen, ob eine E-Mail Spam ist oder nicht. Worin unterscheiden sich nun diese beiden Beispiele? In einem Punkt ganz sicher: in der Toleranz für sogenannte *False Positives/False Negatives.* Dahinter verbirgt sich die Wahrscheinlichkeit, mit der das System fälschlicherweise eine Krebserkrankung bzw. Spam als Output liefert, obwohl die Patientin oder Patient gesund ist bzw. es sich um eine legitime E-Mail handelt (*False Positive*) und die Wahrscheinlichkeit, mit der das System fälschlicherweise eine Probe als „gesund" deklariert bzw. eine E-Mail als „nicht Spam", obwohl eine Krebserkrankung vorliegt bzw. es sich bei der E-Mail um Spam handelt (*False Negative*). Im Gesundheitswesen wollen wir sicherstellen, dass die Einschätzung des Systems eine hohe Zuverlässigkeit hat, schließlich geht es um das Leben eines Menschen. Beim E-Mail-Filter wiederum ist es zwar nicht schön, wenn eine Spam im normalen Postfach landet, solange man aber nicht auf etwaige Phishing-Angriffe eingeht oder Links dieser E-Mail tätigt, stellt dies darüber hinaus keine Bedrohung dar. Weiterhin kann durch gelegentliches Sichten des Spam-Ordners sichergestellt werden, dass nicht versehentlich korrekte Emails dort gelandet sind. Eine inkorrekte Klassifizierung ist also insbesondere bei kritischen Einsätzen möglichst zu vermeiden. Ein tieferer Einblick dazu folgt in Kap. 4.

3.8 Übung

1. Welche Verfahren sind üblich beim Trainieren von KI-Systemen?
   a) Gemeinsames Lehren und Lernen im Team
   b) Überwachtes, unüberwachtes und verstärkendes Lernen
   c) Kooperierendes und selbstständiges Lernen
2. Was ist für das überwachte Lernen zwingend notwendig?
   a) Es ist ausreichend, wenn man viele Daten hat, unabhängig davon, in welcher Form diese vorliegen.
   b) Idealerweise wird der Datensatz aufgeteilt in Trainings- und Testdaten. Die Testdaten werden zum Lernen selbst zwar nicht zwingend benötigt, mit diesen kann man jedoch überprüfen, wie zuverlässig das System den richtigen Output liefert.
   c) Es müssen Daten vorliegen, die über ein zugehöriges Label verfügen.
3. Welche Aussagen über unüberwachtes Lernen sind zutreffend?
   a) Mittels unüberwachtem Lernen können Cluster und Zusammenhänge in großen Datensätzen erkennbar werden, die ansonsten nicht ersichtlich sind.
   b) Es ist einfacher, als beim überwachten Lernen, die Zuverlässigkeit zu überprüfen.
   c) Unüberwachtes Lernen funktioniert mit großen Datensätzen ohne Label.
   d) Es ist schwieriger, die Zuverlässigkeit des Systems zu überprüfen, weil es keinen Testdatensatz gibt.
4. Ist die folgende Aussage korrekt: Beim verstärkenden Lernen lernt das KI-System spielerisch, es wird für korrekte Ausgaben belohnt und für inkorrekte Ausgaben bestraft. Häufig wird verstärkendes Lernen in der Robotik eingesetzt, es wurde aber auch für das Trainieren von Schach oder dem asiatischen Strategiespiel Go genutzt.
   a) Diese Aussage ist korrekt.
   b) Diese Aussage ist nicht korrekt.
5. Künstliche neuronale Netze werden in die folgenden Schichten aufgeteilt:
   a) Input-Schicht, versteckte Schicht (Hidden Layer), Output-Schicht
   b) KNNs werden nicht in Schichten unterteilt
   c) Aufnahme- und Abgabeschicht
   d) Tages- und Nachtschicht
6. Jedes Neuron in einem KNN enthält eine sogenannte Aktivierungsfunktion, diese besteht aus der Summe der gewichteten Ausgabewerte der vorigen Neuronen.
   a) Diese Aussage ist nicht korrekt.
   b) Diese Aussage ist korrekt.

7. Welche dieser Aussagen über SVM sind korrekt?
   a) SVM ist ein Algorithmus, mit dessen Hilfe man Daten klassifizieren kann.
   b) Ziel ist es, dass zwischen Hyperebene und dem nächsten Punkt auf beiden Seiten der maximal mögliche Abstand erreicht wird.
   c) Ziel ist es, dass zwischen Hyperebene und dem nächsten Punkt auf beiden Seiten der minimal mögliche Abstand erreicht wird.
   d) Eine Support Vektor Maschine (SVM) ist eine Maschine im physischen Sinne.
8. Bei der linearen Regression ist der Abstand zwischen jedem realen Datenpunkt und dem zugehörigen berechneten und vorhergesagten Punkt so groß wie möglich.
   a) Diese Aussage ist korrekt.
   b) Diese Aussage ist nicht korrekt.
9. Sie haben sowohl den Algorithmus Support Vector Machine (SVM) als auch die Logistische Regression in diesem Modul kennengelernt. Beide Verfahren unterteilen die Daten in zwei oder mehr Ausprägungen. Worin unterscheiden sich die beiden aber grundsätzlich?
   a) Die Verfahren unterscheiden sich nicht, es hängt von der Laune und den Vorlieben des Data Scientist ab, welches Verfahren genutzt wird.
   b) SVM ist geometrisch motiviert, es wird geschaut, wo welche Datenpunkte liegen. Die logistische Regression hingegen ist durch ein Wahrscheinlichkeitsdenken angetrieben, hier geht es vielmehr darum, mit welcher Wahrscheinlichkeit die abhängige Variable die eine oder eben die andere Ausprägung hat.

3.9 Aufgaben zum eigenen Anwendungsfall

Sie haben im vorigen Abschnitt erarbeitet, welche Daten Sie für Ihr KI-System nutzen möchten. Nun lautet die Frage, in welcher Form das System die Daten prozessieren soll:

- Inwiefern möchten Sie überwachtes, unüberwachtes oder verstärkendes Lernen anwenden? Wägen Sie die Vor- und Nachteile ab und begründen Sie Ihre Entscheidung.
- Wenn Sie überwachtes Lernen nutzen möchten: verfügen Ihre Daten über ein Label? Beschreiben Sie, weshalb Ihre Daten über ein Label verfügen bzw. wie Sie Label generieren möchten, falls es bisher keine gibt.

- Soll ein künstliches neuronales Netz mit den Daten trainiert werden oder ist Ihr Anwendungsfall passender für eine lineare oder logistische Regression? Könnte statt einer logistischen Regression die Möglichkeit der Support Vector Machine eventuell angebrachter sein?
- Erörtern Sie insgesamt, warum Sie sich für ein bestimmtes Verfahren entscheiden. Dieser Kurs hat Ihnen hierfür Grundlagen vorgestellt, an dieser Stelle könnten Sie Ihr Wissen über die spezifischen Techniken im Rahmen weiterführender Literatur vertiefen.

Literatur

Buxmann, P., & Schmidt, H. (2019). Grundlagen der Künstlichen Intelligenz und des Maschinellen Lernens. In P. Buxmann & H. Schmidt (Hrsg.), *Künstliche Intelligenz: Mit Algorithmen zum wirtschaftlichen Erfolg* (S. 3–19). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-57568-0_1.

Ertel, W. (2016). *Grundkurs Künstliche Intelligenz. Eine praxisorientierte Einführung* (4. Aufl.). Springer Vieweg.

Kreutzer, R. T., & Sirrenberg, M. (2019). *Künstliche Intelligenz verstehen. Grundlagen – Use-Cases – Unternehmenseigene KI-Journey* (R. T. Kreutzer & M. Sirrenberg, Hrsg.). Springer Fachmedien. https://doi.org/10.1007/978-3-658-25561-9_1.

Russell, S., & Norvig, P. (2012). *Künstliche Intelligenz* (Bd. 2). Pearson Studium.

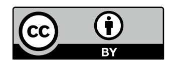
Seegerer, S., Michaeli, T., & Romeike, R. (2020). So lernen Maschinen! *LOG IN 193/194,* S. 27–31.

Weiterführende Literaturhinweise

Kriesel, D. (2005). *Ein kleiner Überblick über Neuronale Netze*. http://www.dkriesel.com/science/neural_networks. Zugegriffen: 15. Okt. 2022.

Vishal, M., & Samer, S. (2017). *Machine learning for humans*. https://www.dropbox.com/s/e38nil1dnl7481q/machine_learning.pdf?dl=0. Zugegriffen: 15. Okt. 2022.

Wittpahl, V. (2019). *Künstliche Intelligenz. Technologie | Anwendung | Gesellschaft*. Springer.


BY

Grundlagen: Output

4

Zusammenfassung

Neben den bereits vorgestellten Inputs und der Funktionsweise sind vor allem die Ergebnisse von KI-Systemen entscheidend. Sie bieten die Grundlage für alle weiteren Verarbeitungen und Handlungen, sei es durch Mensch oder Maschine. Dieses Kapitel macht Unterschiede zwischen den verschiedenen Kategorien von Outputs deutlich, bspw. inwiefern ein System nur Kategorien kennt oder einen kontinuierlichen Wert bereitstellt. Dazu werden verschiedene Arten von KI-Outputs, sowie das Konzept von Metadaten an zwei Fallbeispielen vorgestellt und durchexerziert.

4.1 Einleitung

An dieser Stelle ist schon einiges über KI-Système bekannt (siehe Kap. 2 und 3): Welche Daten werden für diese genutzt? Wie werden diese bearbeitet? Jetzt geht es um den letzten Teil der Einführung, nämlich die Frage: was für Ergebnisse erzeugt ein solches System?

In diesem Kapitel geht es demnach um die Resultate, die KI-Systeme erzeugen. In der Anwendung ist das eine wichtige Fragestellung: denn der Output eines KI-Systems ist die Information, die z. B. zur weiteren Bearbeitung oder Entscheidung an den Menschen übergeben wird. Plant man den falschen Output, können die Nutzenden vielleicht nichts damit anfangen – wenn zum Beispiel ein Bescheid angenommen oder abgelehnt werden muss und das System lediglich eine Prozentzahl darstellt. Oder wenn ein Tool zur Verarbeitung von Texten zurückmeldet, ob ein Text leicht verständlich geschrieben ist, obwohl eigentlich die Anzahl an Wörtern benötigt wird.

M. Heine et al., *Künstliche Intelligenz in öffentlichen Verwaltungen*, Edition eGov-Campus, https://doi.org/10.1007/978-3-658-40101-6_4

Es gibt viele Arten des Outputs, die zu dem Problem, das man lösen will, passen können. Dieses Kapitel soll eine kurze Einführung zu unterschiedlichen Typen geben und dabei helfen, diese anhand von konkreten Beispielen zu verstehen und zuzuordnen.

Als Grundlage werden zunächst zwei Fallbeispiele vorgestellt. Danach werden nach und nach verschiedene Arten von Ergebnissen erklärt – und auf die Beispiele bezogen. Nach jeder dieser kurzen Abschnitte gibt es zudem Aufgaben, die Sie dabei unterstützen sollen, sich tiefer mit den dargestellten Themen zu beschäftigen.

4.2 Fallbeispiele

Am Ende eines Arbeitsprozesses können verschiedene Formen von Ergebnissen stehen. Das kann zum Beispiel die Entscheidung darüber sein, ob ein Antrag bewilligt wird oder nicht. Genauso gut kann es aber auch eine Prognose sein, z. B. wieviele PKW eine bestimmte Kreuzung nutzen werden. Das Ergebnis könnte weiterhin z. B. eine E-Mail oder ein Brief mit der Bitte um eine Auskunft sein.

Diese verschiedenen Arten von Ergebnissen verbindet eines: es wurden Informationen genutzt, damit sie zustande kommen. Möchte man die Nutzungsfrequenz einer Kreuzung prognostizieren, können dazu historische Daten über die Nutzung eingesetzt werden. Es können Vergleichsdaten von anderen Kreuzungen genutzt werden, um ein besseres Ergebnis zu erzielen.

In diesem Modul soll es darum gehen, welche Ergebnisse ein intelligentes System eigentlich erzeugen kann – und wo es insbesondere anderen Systemen gegenüber überlegen sein kann. Dazu werden wir uns als Beispiel zwei Systeme anschauen, die als Fallbeispiele dienen sollen:

1. Schreibfix – dieses System analysiert geschriebene Texte, z. B. hinsichtlich ihrer Verständlichkeit, der generellen Stimmung und der Ähnlichkeit mit anderen Texten.
2. Memoriali – dieses System verarbeitet Anträge auf Änderungsgenehmigung für denkmalgeschützte Gebäude und gibt Empfehlungen für den Bescheid aus.

In den beiden folgenden Texten werden die beiden Systeme kurz beschrieben. Über Abschn. 4.3 finden Sie Fragen zu den Texten.

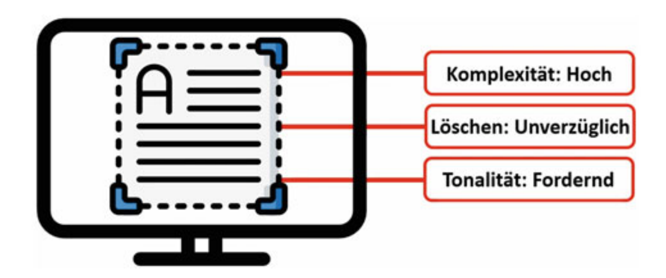


Abb. 4.1 Schreibfix

4.2.1 SchreibFix

Mithilfe von SchreibFix schreibt Thomas eine E-Mail an seine Vorgesetzte, Mika. Thomas kennt Mika erst seit Kurzem. Er möchte in seiner E-Mail freundlich klingen, aber – da Mika bereits einige E-Mails ignoriert hat – auch selbstbewusst auftreten. Oft drückt er sich zu kompliziert aus. War das vielleicht der Grund?

In dieser E-Mail möchte er sie darauf hinweisen, dass das Meeting mit dem gesamten Bereich noch nicht vorbereitet ist, er aber ein paar Ideen hat. Er bereitet die E-Mail vor und hofft, dass ihm so ein guter Einstieg mit der neuen Sachgebietsleiterin gelingt. Er benötigt dazu auch noch eine schnelle Antwort. Hat sich da in seine E-Mail ein „unverzüglich“ eingeschlichen?

Zum Glück ist er nicht allein. Das Tool „SchreibFix“ (Abb. 4.1) erkennt, dass ein Text kompliziert ist, an einigen Stellen länger als notwendig – und vor allem sehr fordernd. Thomas ändert seinen Text. Kein „prophylaktisch“ und das „unverzüglich“ fliegt raus. So klappt es mit dem E-Mail-Schreiben, dank SchreibFix!

4.2.2 Memoriali

Sie arbeiten im Herzen einer Altstadt – wunderbare, historische Gebäude, altehrwürdige Fassaden – Geschichte zum Anfassen. Das ist wirklich etwas Besonderes – und das soll so bleiben.

Deswegen gibt es Verordnungen zum Schutz solcher Gebäude. Dies ist sehr wichtig, die Verordnungen müssen jedoch manchmal außer Kraft gesetzt werden. Zum Beispiel, wenn Fenster dringend erneuert werden müssen, damit das Gebäude nicht von innen heraus schimmelt oder überaus ineffizient beheizt wird.

Auf Antrag kann der Denkmalschutz zur Erneuerung des Gebäudes außer Kraft gesetzt werden. Diese Anträge sind häufig kompliziert – sowohl für Sie als auch für die Antragstellenden. Zwischen mehreren Zielen muss abgewogen werden: einerseits soll der historische Wert des Gebäudes erhalten werden – andererseits gibt es gesundheitliche oder ökonomische Herausforderungen.

Um diese beiden Ziele miteinander überein zu bringen, gilt es in dem Antrag zu zeigen, weshalb eine Erneuerung unumgänglich ist – und auch, wie diese stattfinden kann, ohne zu große Veränderungen hervorzurufen. Dafür ist eine Reihe an Unterlagen, Planungen oder sogar Sachverständigenberichten einzureichen. Und anschließend zu prüfen.

Memoriali vereinfacht diesen Prozess – für Antragstellende und Sachbearbeitende. Wer bauliche Veränderungen vornehmen muss, wird durch den Prozess der Antragstellung Schritt für Schritt begleitet. Verschiedene Beantragungsziele werden erkannt und bei der Einreichung relevanter Unterlagen wird viel Unterstützung geboten. Nach erfolgter Beantragung berechnet das Programm durch intelligente Regeln und einen Vergleich zu ähnlichen Verfahren mögliche Bescheide sowie Änderungen, die im Konzept vor einer Bewilligung notwendig sind.

▶ Memoriali unterstützt alle Parteien in diesem wichtigen, aber komplexen Prozess.

Übung

Zur Vorbereitung auf den kommenden Abschnitt, bearbeiten Sie die folgenden Fragen für beide Systeme. Achten Sie besonders auf die Unterschiede zwischen den Systemen.

1. Welche Aufgaben hat das System in dem Fallbeispiel zu erledigen?
2. Gibt es eindeutig richtige und falsche Ergebnisse bezüglich der Aufgaben, die das System übernimmt?
3. Würden Sie das System als intelligent beschreiben? Was spricht dafür, was dagegen?
4. Kann das System unabhängig von einem Menschen arbeiten? Falls nein, welche Kooperation mit Menschen ist erforderlich?

4.3 Kategorien

Ergebnisse von KI-Systemen können sehr unterschiedlich aussehen. Das hängt damit zusammen, dass unter KI auch sehr viele unterschiedliche, statistische Methoden zusammengefasst sind.

Ein sehr prominentes Beispiel ist die Sortierung von eingegangenen Informationen in verschiedene Kategorien. Diese Kategorien werden – im Regelfall – vor dem Training der KI festgelegt und können danach als Ergebnisse ausgegeben werden. Manchmal entwickeln KI-Systeme auch eigene Sortierungen für Daten, siehe Abschn. 4.4 zu Pattern Matching.

Mit Blick auf die beschriebenen Fallbeispielen zeigt sich folgendes Bild (siehe Abb. 4.2):

- SchreibFix analysiert geschriebene Texte, z. B. hinsichtlich ihrer Verständlichkeit, der generellen Stimmung und der Ähnlichkeit mit anderen Texten.
- Wir wollen uns in diesem Beispiel auf das Thema „Stimmung" konzentrieren.
- Stellen Sie sich vor, Sie formulieren eine E-Mail an eine Bürgerin. Der Sachverhalt ist dringlich und Sie benötigen eine rasche Antwort. Gleichzeitig wollen Sie aber auch herzlich klingen, da die Bürgerin sich in einer schweren Lebenslage befindet.

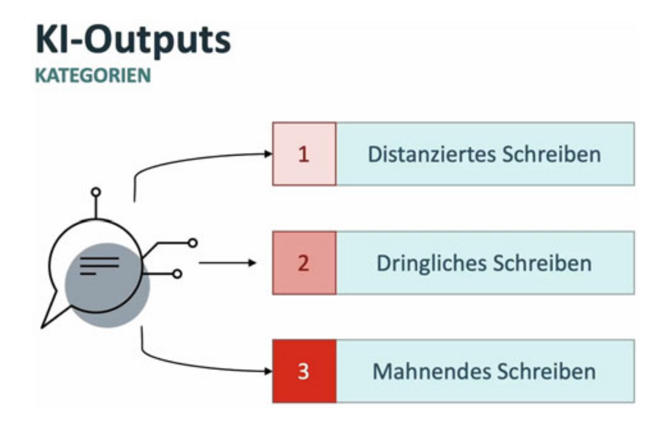


Abb. 4.2 Kategorien SchreibFix

- Nachdem Sie Ihre E-Mail entworfen haben, nutzen Sie SchreibFix, um die Stimmung Ihres Textes zu überprüfen. Das Ergebnis ist die Kategorie „Dringliches Schreiben“. Sie sind zufrieden.
- Sie können auch noch überprüfen, welche anderen Kategorien für das System naheliegend waren: „Distanziertes Schreiben“ und „Mahnendes Schreiben“ stehen dort ganz oben. Das entspricht nicht Ihren Vorstellungen und Sie passen den Text noch einmal an.

Reflektieren Sie das Fallbeispiel mit der folgenden Frage Durch wen oder was wird im Falle von SchreibFix definiert, was ein „Distanziertes Schreiben“ ist?

Die „Stimmung“ in dem Beispiel ist eine Kategorie. Sie zeichnet sich dadurch aus, dass es keine Reihenfolge zwischen Kategorien gibt. „Distanziertes Schreiben“ ist nicht besser oder schlechter als „Mahnendes Schreiben“. Ob es passt oder nicht, hängt von Ihrem Anwendungsfall ab.

Dass es unterschiedliche Kategorien gibt, sorgt dafür, dass manche von ihnen mehr oder weniger zutreffen können als andere. Bei kategorialen Einstufungen durch KI-Systeme kann es daher wichtig sein, nicht nur die wahrscheinlichste Kategorie zu betrachten, sondern auch Kategorien, die mit einer niedrigeren Priorität ausgewählt wurden. Einerseits können diese weitere Hinweise enthalten, andererseits kann durch die Eigenheiten Ihres Schreibstils vielleicht auch die zweitstärkste Kategorie aussagekräftig sein.

Es gibt viele Gründe, bei der Einführung automatisierter Systeme nicht nur das Endergebnis eines Prozesses zu zeigen, sondern auch Meta-Informationen oder wahrscheinliche Alternativergebnisse. Im Kap. 8 zu erklärbarer KI werden noch verschiedene Beispiele dazu diskutiert.

Kategorien können aber auch bei Bildern gebildet werden. Die Bilderkennung ist ein klassisches Beispiel für den Einsatz von Künstlicher Intelligenz. Dabei können Bilder zum Beispiel in zwei einfache Kategorien unterteilt werden („Hund“/ „kein Hund“). Allerdings könnte es auch komplexere Kategorien geben, wie z. B. die Bestimmung von Pflanzen- oder Tierarten anhand von Bildaufnahmen. Diese können bei der Arbeit von Sachverständigen genutzt werden.

Den vielleicht stärksten Einfluss auf die Verwaltung dürfte KI aber in der automatisierten Vorsortierung schriftlicher Anfragen haben. Hier können durch KI-Systeme bestimmte Fachbereiche oder sogar Personen zugeordnet werden. Solange einem KI-System im Training unterschiedliche Kategorien zu Trainingsdaten geliefert werden, sind der Innovation keine Grenzen gesetzt. Auch Personen, Bearbeitungsfristen oder Zuständigkeitsbereiche sind Kategorien, in die Texte, Bilder, Videos oder andere Daten eingeordnet werden können.

Das Fallbeispiel zeigt aber auch, dass die Art und Weise, wie Kategorien zugewiesen werden, ein statistischer Prozess ist, der von den Daten, mit denen das System trainiert wurde, abhängt. Es kann also in so einem Fall sein, dass der spezifische Schreibstil nicht genau von SchreibFix erfasst werden kann. Woran könnte das liegen?

Vielleicht werden im Text ungewöhnliche Begriffe verwendet oder Worte, die nur in einem bestimmten Fachbereich genutzt werden. Das KI-System kennt diese Begriffe dann nicht aus dem Training und wäre damit nicht in der Lage, den Text so gut zu klassifizieren, wie andere Texte.

Übung
In diesem Abschnitt haben Sie etwas über Kategorien als möglicher Output von KI-Systemen gelernt. Versuchen Sie sich an den folgenden Aufgaben, bevor Sie fortfahren:

1. Für die Stadtplanung muss beim Bau einer Brücke die Statik überprüft werden. Ein Kollege schlägt vor, die Informationen zum Bau einem KI-System mit den Kategorien „sicher“ und „nicht sicher“ zu geben. Wie beurteilen Sie dieses Vorhaben?
2. Sie erstellen eine wiederholung Einladung zu einem Gespräch für einen Bürger. SchreibFix gibt Ihnen als Einstufung der „Stimmung“ in Ihrem Text „Freundliches Schreiben“ an. Sie möchten gerne erreichen, dass daraus ein „Mahnendes Schreiben“ wird. Was könnten Sie überprüfen?
3. Ein neues KI-System zur automatischen Zuweisung von Schriftverkehr soll eingerichtet werden. Wie gehen Sie vor, um entsprechende Kategorien zu entwickeln? Welche Herausforderungen könnten dabei entstehen?
4. Planen Sie nun die Kategorien, die in einem eigenen Prozess in Ihrem Beruf relevant sind. Beschreiben Sie, welche Kategorien Sie definieren würden, welche Vorteile daraus entstehen würden und welche Daten für das Training vorhanden sein müssten.

4.4 Pattern Matching

Die Zuordnung von Daten zu Kategorien haben wir bisher vor allem mit statischen Daten wie einem Bild oder einem Text beleuchtet. Allerdings kann es auch vorkommen, dass Daten über einen bestimmten Zeitraum beobachtet werden müssen. Nehmen wir zum Beispiel die Anzahl an Krankheitsfällen in einer

Organisation. Wenn bemerkt wird, dass vermehrt Personen krank werden – und das über einen längeren Zeitraum – und eventuell auch noch der Herbst beginnt, kann daraus eine Schlussfolgerung gezogen werden: die Erkältungswelle rollt.

Um das zu tun, wurden wiederkehrende Informationen genutzt – die Anzahl an Kranken an verschiedenen Tagen. Da in der Vergangenheit bereits beobachtet werden konnte, dass unter diesen Umständen eine Erkältungswelle besteht, kann nun ein entsprechender Rückschluss gezogen werden. Am Ende dieses Kapitels geht es auch um Möglichkeiten, sogar zukünftige Entwicklungen abschätzen zu können. Allerdings sind nicht nur numerische Daten die Grundlage dieser Form der Mustererkennung, auch Pattern Matching genannt (vgl. Russel & Norvig, 2012, S. 400). Generell ist das Pattern Matching eng verwandt mit der Sortierung in Kategorien. Wieso unterscheidet sich dies von Kategorien? Für ein Pattern Matching werden stets mehrere, z. B. zeitlich getrennte Datenpunkte, benötigt. Bei Kategorien wie einer Bildanalyse ist dies nicht notwendig. Möchten Sie zum Beispiel die Strategie eines Spielers beim Schach erkennen, ist es mitunter notwendig, mehrere Züge zu beobachten, um eine zuverlässige Aussage zu treffen. Eine Momentaufnahme würde dafür nicht ausreichen – könnte aber helfen um z. B. zu bewerten, welcher Spieler oder welche Spielerin gerade größere Chancen auf den Sieg hat.

Das Pattern Matching hat außerdem noch eine andere, wichtige Komponente. Durch die wiederholte Beobachtung von Ereignissen und Daten können Zusammenhänge, also Muster – die Patterns – identifiziert werden. Es muss also vorher nicht in einem Datensatz für das Training bereits ein „Label“ entwickelt werden. Die Muster können dadurch erkannt werden, dass bestimmte Abfolgen von Informationen mit einer hohen, statistischen Wahrscheinlichkeit nacheinander auftreten.

Immer wenn also Verläufe von Daten zur Verfügung stehen, die abstrakteren Mustern zugeordnet werden können, ist es sinnvoll Pattern Matching als Verfahren einzusetzen. Dies könnten zum Beispiel Kontostände sein, die sich über die Zeit verändern – und ggf. Rückschlüsse darauf zulassen, ob ein Unternehmen von der Kleinunternehmerregelung betroffen sein kann oder nicht. Ein anderes Beispiel sind Daten zum Aufkommen von Wildtieren, die eine Aussage darüber ermöglichen, ob die Tiere bereits in der Brunftphase sind – und so eine präzise Steuerung von Schutzphasen zulassen.

Fallbeispiel Memoriali

Schauen wir uns das Beispiel von Memoriali einmal an:

Alberta Fischer lebt in einem älteren Haus, welches unter Denkmalschutz steht. Die Fenster sind aus Holz und trotz einer guten Pflege ist dieses inzwischen verrottet. Sie entschließt sich, die Fenster erneuern zu lassen – der Fensterbauer macht sie darauf aufmerksam, dass es notwendig sei, einen Antrag bei der Stadtverwaltung zu stellen. Sie nutzt dafür das Angebot der Stadt, dies online via Memoriali zu erledigen.

Zunächst wird sie im Programm gebeten, einige Daten einzutragen: das Alter und den Standort des Hauses, das geplante Vorhaben etc. Kurz darauf erscheint eine Pop-Up-Benachrichtigung: „Wir haben 17 ähnliche Vorgänge für Häuser in Ihrer Umgebung gefunden. Möchten Sie angezeigt bekommen, welche Unterlagen Sie für dieses Vorhaben benötigen?“. Erfreut klickt Angelika auf „Ja“ und wird vom Programm danach zielstrebig unterstützt.

In diesem Beispiel fand eine Form des Pattern Matchings statt – die Art und Weise, auf die Angelika Fischer mit dem System interagiert hat, war die Grundlage dafür, dass das System erkennen konnte, was sie plant. Dadurch konnte ein spezifischer Vorgang vorgeschlagen werden. Diese Vorgehensweise kann manche Nutzende aber auch erschrecken, weil sie sich beobachtet fühlen. Deshalb sollte vor der Nutzung auf die Funktionsweise des Programms hingewiesen werden.◄

Self-Organizing Maps

Ein weiterer Anwendungsfall für Pattern Matching sind die sogenannten Selbstorganisierenden Karten (vgl. Carrasco & Brunner, 2014). Diese Technik des maschinellen Lernens kann genutzt werden, um zweidimensionale, topologische Karten höherer und komplexerer Datensätze zu erzeugen – so beispielsweise zum semantischen Clustering unterschiedlicher Eigenschaften von Altbauten, wie Grad der Wärmedämmung, Ästhetik des Gebäudes, Verfallsgrad oder Ähnliches. Angelehnt an die Art und Weise, wie unser Nervensystem Informationen verarbeitet, wird hier Nähe genutzt, um Ähnlichkeit zwischen verschiedenen Daten anzuzeigen. Dabei landen semantisch ähnliche Datenpunkte auf der Karte näher beieinander und bilden so nach und nach klar abgrenzbare Bereiche für ihre jeweiligen Cluster, bspw. in Form kleiner Insel- oder Landmassen.

Übung
In diesem Abschnitt haben Sie etwas über das Pattern Matching als möglicher Anwendung für intelligente Systeme erfahren.

1. Sie erheben regelmäßig Daten zur aktuellen Belastung eines Sees mit Algen. Dieser muss regelmäßig und rechtzeitig gesperrt werden, dies hängt aber auch stark vom Wetter ab. Wie könnten Sie Pattern-Matching-Verfahren verwenden, um dies zu ermöglichen?
2. Sie erhalten einen anonymen Hinweis mit dem Auszug der heutigen Kontodaten einer Person; Sie wollen prüfen, ob hier Geldwäsche betrieben worden sein könnte. Inwiefern können Sie diese Aufgabe mit Pattern Matching lösen?
3. Im Rahmen einer Erweiterung von Memoriali werden in regelmäßigen Abständen die Dichtigkeiten der Fenster in Ihrem Verwaltungsgebäude überprüft. Wie könnte Pattern Matching Ihnen helfen, mit diesen Daten effizient zu verfahren?
4. Beziehen Sie sich nun auf einen Sachverhalt aus Ihrem aktuellen Beruf oder Projekt: Welche Daten liegen im Rahmen einer Zeitreihe vor? Welche Einsichten könnte die Beobachtung von Mustern in diesen Zeitreihen bereitstellen?

4.5 Numerische Prädiktion

Machine Learning kann genutzt werden, um verschiedenste Arten von Outputs zu generieren. Eine der vielleicht Wichtigsten ist dabei die Erzeugung von sogenannten numerischen Prädiktionen (siehe Abb. 4.3), also der Ausgabe von Zahlenwerten, um Vorhersagen über einen Sachverhalt zu ermöglichen.

Im weitesten Sinne können als numerische Prädiktionen alle Outputs verstanden werden, die Zahlen als Ergebnis haben. Dies ist im Sinne von kontinuierlichen Zahlen zu verstehen, die in mathematischen Bezügen zueinanderstehen. Es geht also, anders als bei den zuvor vorgestellten Kategorien, nicht um gleichwertige Unterteilungen mit verschiedenen Charakteristiken, sondern darum, eine mathematische Funktion zu finden, die zu einem gegebenen Datensatz passt.

Eingangswerte können allerdings durchaus auch als Kategorien vorliegen.

Anschauliche Beispiele für verschiedene Outputs dieser Art sind etwa folgende:

- Die durch ein System errechneten Rentenbeiträge, die einer Person im Alter zur Verfügung stehen. Als Grundlage des Datensets dienen hier die von

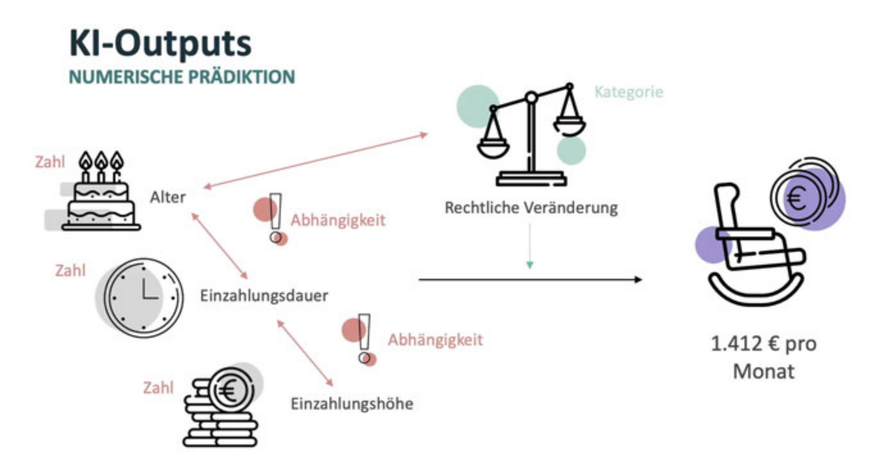


Abb. 4.3 Numerische Prädiktion Rente

der jeweiligen Person eingezahlten Beiträge. Diese werden über den Einzahlungszeitraum hinweg gesammelt und über eine mathematische Formel in sogenannte Rentenpunkte umgerechnet. Ein mit dieser Art Input trainierter Algorithmus könnte jetzt auf Grundlage verschiedener Variablen eine numerische Prädiktion treffen, bspw. wie hoch oder niedrig die auszuzahlende Rente aktuell aussieht. Diese Prädiktion kann dann als Grundlage genutzt werden, um zu beraten oder Anpassungen an der Finanzplanung der Person vorzunehmen. Es handelt sich hierbei klar um eine numerische Prädiktion, da das Ergebnis in einer mathematischen Methode besteht, die nicht nur unterscheidbar von anderen Kategorien ist, sondern eben auch bewertende Aussagen möglich macht (Rente ist höher oder niedriger).

- Ein weiteres Beispiel wäre die Berechnung der Steuerlasten einzelner Bürgerinnen und Bürger: Stellen Sie sich vor, dass Sie zur Berechnung der Steuerlast, die eine Person Zeit ihres Lebens zu tragen hat, ihre Informationen und Finanzdaten in ein System eintragen. Das Ergebnis dieser komplexen, durch den Algorithmus Ihres Hilfstools erfolgten Berechnungen wird eine Kennzahl sein, die eine Aussage darüber ermöglicht, wie die individuelle Steuerlast dieser Person in den nächsten Jahren genau aussehen wird. Auf Grundlage dieses Ergebnisses können Sie zum einen beurteilen, ob diese Last

relativ gesehen hoch oder niedrig sein wird. Sie können sie aber auch direkt mit den Werten anderer Personen vergleichen und so ein Urteil tätigen.

In Bezug auf das präsentierte Fallbeispiel der Altbauten könnte eine mögliche numerische Prädiktion in der Erechnung von durch verschiedene Faktoren beeinflusste Immobilienwerte oder die Kalkulation der Dauer von Umbaumaßnahmen bestehen.

Ein weiterer Aspekt, den es in Bezug auf numerische Prädiktion zu beachten gilt, ist ihr Wert für Machine-Learning-Prozesse wie bspw. Reinforcement Learning (vgl. Abschn. 3.4). Wichtigstes Element des Reinforcement Learnings ist die Idee, dass der lernende Algorithmus nicht nach jedem Datenpunkt Feedback bekommt, sondern erst nach dem eine bestimmte Anzahl an Schritten erfolgt ist. Dies ist besonders bei spielerischen Kontexten der Fall, bei denen nach einer wie auch immer abgegrenzten Runde eine gewisse Punktzahl erreicht werden muss. Bleibt der Algorithmus unter dieser Zahl, erhält er auch keine Belohnung. Um seine Strategie und somit sich selbst zu verbessern, sind mehrere Durchläufe nötig. Numerische Prädiktion erlaubt ein Feedback, das nicht nur eine klar beobachtbare Richtung hat (höher vs. niedriger), sondern zusätzlich auch das Ausmaß der Abweichung beobachtbar macht (5 Punkte von 10 benötigen, um Verstärkung zu erhalten).

Auch im Sinne dieser Lernmethode ist es wichtig darauf zu achten, dass Algorithmen, die für Numerische Prädiktion eingesetzt werden sollen, sehr „saubere" Datensätze für ihr Training benötigen, da starke Ausreißer die Qualität der Beurteilung verringern und die Modelle des Algorithmus verwirren können. Man stelle sich exemplarisch vor, dass für die Berechnung der Rentenpunkte Zahlenwerte eingespeist worden wären, die um ein vielfaches höher oder niedriger waren als „realistische Werte".

Übung

In diesem Abschnitt haben Sie etwas über Numerische Prädiktion als Aufgabe für intelligente Systeme erfahren. Bitte bearbeiten Sie die folgenden Aufgaben, um fortzufahren:

1. Ein neues System zur Berechnung von Bußgeldern bei Verstößen gegen die Straßenverkehrsordnung soll genutzt werden. Würden Sie ein solches System mittels neuronalen Lernen trainieren? Welche Argumente könnten dafür oder dagegen sprechen?
2. Eine benachbarte Kommune hat ein System eingeführt, welches automatisiert Anträge auf Einstufung des Schwerbehinderungsgrads bearbeitet. Dafür wurde

ein System zur Kategorienbildung benutzt. Ihr Kollege schlägt vor, eher auf numerische Prädiktion zu setzen. Welche Argumente für die jeweilige Position gibt es?
3. In Memoriali wurde auch eine Schätzung des Preises für die Umbaumaßnahmen integriert. Allerdings gibt diese keinen genauen Wert (2057,53 €), sondern einen groben Wert an (zwischen 2000 und 2300 €). Handelt es sich hierbei um eine numerische Prädiktion?

4.6 Synthetische Ergebnisse

Verschiedene Outputs von Systemen, die Machine Learning nutzen, wurde in den vorigen Abschnitten bereits vorgestellt. Eine in unserem Alltag häufig vorkommende, aber oft übersehene Variante sind dabei Systeme, die „synthetisierte" oder „simulierte" Ergebnisse erzeugen. Synthetische Ergebnisse sind in Form elektronischer Musik oder in Videospielen wie Minecraft erlebbar und zwar durch automatisch generierte Musikelemente oder prozedural erstellte Spielwelten.

Prozedural generiert bedeutet dabei, dass Inhalte nach bestimmten Regeln generiert werden. In simulierten Welten wird so z. B. geregelt, in welchem Verhältnis Land und Wasser zueinanderstehen oder dass Bäume nur auf festem Untergrund wachsen können und nicht z. B. im Wasser. Der Anwendungsraum dieser Technologien ist allerdings deutlich größer: auch natürliche Prozesse können z. B. in simulierten Umwelten getestet werden – die Prüfung des Luftwiderstandes von Fahrzeugen ist dabei genauso relevant wie z. B. die Erosion von Böden nach der Rodung bestimmter Landabschnitte. Sie können so z. B. überprüfen, wie bestimmte Regeln sich auf unterschiedliche – prozedural generierte – Gegebenheiten auswirken würden.

Ein weiteres, sich immer weit verbreitendes, Beispiel sind DeepFake-Videos. Also Videos, die so digital manipuliert worden sind, dass es aussieht, als ob Personen, z. B. prominente Politiker- oder Künstler/innen, in ihnen auftauchen, die nie im ursprünglichen Video mitgespielt haben. Ein mit DeepFake erzeugtes Video von Tom Cruise, welches auf der Videoplattform TikTok kursierte sowie die Einbindung der eigentlich verstorbenen Schauspielerin Carry Fischer in ihrer Rolle als Prinzessin Leia in den jüngsten Star Wars Filmen, sind populäre Beispiele.

Alle präsentierten Beispiele teilen sich dabei eine technische Grundlage bzw. einen der Technologie zugrunde liegenden Prozess: Bei der Erstellung von

synthetischen Ergebnissen wird zunächst auf Grundlage struktureller Zusammenhänge im Beispielmaterial Regelwissen aufgebaut, auf dessen Grundlage die spätere Synthetisierung erfolgen kann. Wie in Kap. 2 bereits ausgeführt, kann dies durch die Nutzung von Trainingsdaten erfolgen – die Regeln können aber auch explizit gegeben werden.

Im Falle von DeepFakes könnten dies Regeln darüber sein, wie sich die Mimik einer prominenten Person verändert, wenn diese eine bestimmte Handlung wie Lächeln o. ä. vollzieht. In diesem Beispiel wären Trainingsdaten notwendig. In einem nächsten Schritt kann dann die Synthetisierung erfolgen, bei der neue Inhalte erzeugt werden, die den zuvor festgelegten Regeln und Strukturen folgen, sodass am Ende eine synthetische, aber trotzdem authentische Entität erschaffen worden ist.

Welche genaue Form diese annimmt, kann, wie die Beispiele gezeigt haben, vielseitig sein. Wichtig ist die Abfolge von Abstraktion von Regeln, z. B. aus Beispielmaterial, und dann Synthetisierung auf Grundlage dieser Regelstrukturen.

Diese Regeln werden dann auf einen Ausgangswert angewendet. Im Falle prozedural generierter Daten spricht man von einem „Seed“, also dem Samen, der die Grundlage für z. B. prozedural generierte Welten oder auch Musik ist.

Bild- oder Videomaterialen, die via DeepFake oder ähnlichen Systemen verändert wurden, können eine große Herausforderung im Bereich des Datenschutzes oder Urheberrechts darstellen. Auch das Persönlichkeitsrecht könnte davon betroffen sein – mehr dazu im Rechtsteil (Kap. 11).

Beziehen wir diese Idee auf die bereits vorgestellten Fallbeispiele, wird die Variabilität dieses Outputs noch einmal deutlicher:

Ein synthetisches Ergebnis im Falle der Nutzung von SchreibFix wäre beispielsweise eine vollständig synthetisch erzeugte E-Mail. Sie als bearbeitende Person geben dabei klassische Faktoren, wie den Titel der E-Mail, den Adressaten etc., aber auch weitreichendere Informationen, wie das Ziel der E-Mail, z. B. eine Mahnung, und Ihre E-Mail-Historie in das Tool ein. Auf Grundlage dieser Faktoren erstellt die KI dann einen für diesen Vorgang spezifischen Vorschlag. Anders als bei klassischen, lediglich auf vorgefertigten Regeln basierenden Systemen werden hier nicht einfach bereits bestehende Textbausteine konzipiert, sondern erlernte Regeln auf Grundlage Ihrer Eingaben abgeleitet und zur Vorlagen-Erstellung verwendet. So ist es auch möglich, dass sich ein synthetisiertes Produkt auf Ihren eigenen Schreibstil hin anpasst, da das System gelernt hat, wie Sie verschiedene Satzelemente, bspw. Nebensätze, bilden.

Übung
In diesem Kapitel wurden Systeme behandelt, die synthetische Ergebnisse erzeugen. Folgende Aufgaben greifen dies auf:

1. Ein guter Seed für z. B. die Erzeugung einer Mail in SchreibFix sollte möglichst eindeutig sein. Wenn der Seed z. B. nur „Erinnerung" lautet, könnten Informationen fehlen und sehr generische Regeln angewendet werden. Welche Anleitung würden Sie Sachbearbeitenden geben, wenn Sie Seeds für Mails erstellen?
2. Die Generierung von synthetischen Daten steht im Bereich der Validierung von Beweismaterial ein großes Risiko dar. Wie könnten synthetische Daten Sie in Ihrer persönlichen Arbeit vor Herausforderungen stellen? Wie können Sie sich darauf vorbereiten?
3. Bei der Planung des Busverkehrs in der Stadt Schlauheim wird über verschiedene Modelle diskutiert: verschiedene Taktungen, Sammeltaxen in der Nacht oder auch On-Demand-Verkehr. Wie könnte die Erstellung synthetischer Situationen bzw. Simulationen helfen, die Wirksamkeit unterschiedlicher Regelungen festzustellen?

4.7 Forecasting

Ein weiterer wichtiger Output bei der Verwendung von Machine Learning sind Forecastings, also Ergebnisse, die als Vorhersage für bestimmte Ereignisse genutzt werden können. Dabei ist es wichtig zu verstehen, dass es sich bei Forecastings nicht einfach um einen Sammelbegriff für jede Form von Prädiktionen wie bspw. die Numerische Prädiktionen handelt. Der Begriff Forecasting beschreibt vielmehr einen konkreten Prozess, bei dem Daten aus der Vergangenheit und Gegenwart genutzt werden, um durchgängige Verläufe oder Trends für die Zukunft zu antizipieren. Vereinfacht gesagt werden Vorhersagen über mögliche Verläufe getroffen. Standardmäßig wird es für Zeitreihen benutzt.

Exemplarische Beispiele aus dem Alltag sind Trendvorhersagen, wie sie in Wetterberichten genutzt werden, bei denen Wetterdaten aus der Vergangenheit beobachtet werden, um Aussagen über ähnliche Verhältnisse in der Zukunft zu entwickeln. Auch die Prädiktion von Verkehrsaufkommen, wie es viele Apps für die Planung von Routen nutzen, kann als Forecasting bezeichnet werden. Auch die Entwicklung von Absatzzahlen spielt z. B. in der Logistik eine wichtige Rolle.

Für die bereits bekannten Fallbeispiele würde ein exemplarischer Forecasting-Output wie folgt aussehen:

- **Fall I SchreibFix:** Hier könnten Sie auf Grundlage von historischen Daten versuchen herauszufinden, ob es Zeiten im Jahr gibt, an denen die Arbeitslast durch die ankommende E-Mail-Last besonders hoch ist und wie sich das E-Mail-Aufkommen pro Stunde an einzelnen Tagen darstellt. Was Ihnen nicht nur Einblick in Strukturen bietet, die Sie sonst eventuell nur subjektiv wahrgenommen hätten („Im Januar scheint es immer mehr Arbeit zu geben."), sondern es Ihnen auch erlaubt, zusätzlich eine gezieltere Ressourcenplanung für Ihre einzelnen Arbeitstage vorzunehmen.
- **Fall II Memoriali:** Im Sinne der Instandhaltung von Altbauten könnte hier die Verwitterung von Häusern als Ziel des Forecastings begriffen werden. Aus Daten über den Verfall anderer Immobilien in der Vergangenheit, beeinflusst durch Faktoren wie Wettereinflüsse oder Bausubstanzen, kann der Algorithmus Vorhersagen über den zukünftigen Verlauf der Verwitterung einzelner Bauteile treffen.

Forecasting ist dabei mehr, als nur die bloße Voraussage eines einzelnen möglichen Trends. Vielmehr geht es darum, Interaktionen zwischen verschiedenen Prädiktionen sichtbar zu machen. Dies ermöglicht eine präzise Steuerung von z. B. logistischen Ressourcen, da ein Forecasting es absehbar macht, wie diese optimal verteilt werden können.

Rückblickend auf das Fallbeispiel I SchreibFix kann die Voraussage einer Auslastung durch zu hohes E-Mail-Aufkommen als Interaktion bezeichnet werden. Der aufgezeigte Trend zeigt zunächst eine hohe Wahrscheinlichkeit für ein erhöhtes E-Mail-Aufkommen an, um dann die Ressourcenprädiktion „Sind ausreichend Arbeitskräfte verfügbar. Ja oder nein?" zu ermöglichen. Das Forecasting ermöglicht also Einblicke in verschiedene zukünftige Interaktionen.

Übung

In diesem Kapitel wurden das Thema Forecasting behandelt. Die folgenden Aufgaben sollten bearbeitet werden, bevor man fortfährt:

1. Sie haben Messdaten über das Verkehrsaufkommen an einer Kreuzung an Werktagen. Für eine kommende Demo möchten Sie auch die Verkehrslage für einen Samstag beurteilen und darauf auf Forecasting zurückgreifen – Ihnen liegen schließlich mehre Monate an Daten vor. Welche Nachteile erwarten Sie? Welche Aussagen lassen sich ggf. trotzdem treffen?

2. Sie arbeiten im Bereich Arbeitsmedizin und Impfschutz und wollen SchreibFix nutzen und abschätzen, wieviel Anfragen Sie im kommenden Monat erhalten – oft steigt die Anzahl an Anfragen in Ihrer Behörde, wenn die Menschen in den Urlaub gehen. Ein Kollege entgegnet, dass „Mails nicht wie das Wetter sind.“ Wie schätzen Sie diese Aussage ein?
3. Pattern Matching und Forecasting haben sehr große Überlappungen. Welche Beispiele gibt es für Pattern-Matching-Aufgaben, die sich nicht auf Forecasting übertragen lassen und umgekehrt?

4.8 Metadaten von Ergebnissen

Nachdem nun verschiedene Output-Typen für Machine-Learning-Algorithmen vorgestellt wurden, folgt nun ein Blick auf Informationen, die diese Outputs begleiten bzw. aus ihnen generiert werden können, sogenannte Metadaten.

Mit Metadaten kann beschrieben werden, wie gut ein durch Machine Learning entwickeltes Modell für ein Problem angepasst ist (vgl. Sokolova et al., 2006). Diese Daten helfen also zu verstehen, ob es sinnvoll ist, ein Modell einzusetzen oder ob es z. B. mit anderen Daten erneut trainiert werden muss.

Eine wichtige Metrik für die Beurteilung der Qualität von Algorithmen ist dabei die *Accuracy* (siehe Abb. 4.4).

Damit gemeint ist die Qualität, mit der ein Algorithmus richtige Entscheidungen bei der Zuordnung trifft. Versteht man den Algorithmus als System, das in Abhängigkeit von den ihm gegebenen Inputs eine Entscheidung darüber fällt, ob der Input positiv oder negativ ist, dann ist die Accuracy ein Maß dafür, wie oft der Algorithmus richtige positive und richtige negative Inputs identifizieren konnte.

Ein mögliches Anwendungsbeispiel wäre ein Algorithmus, der trainiert werden soll, um unerwünschte Bilder oder Videos (bspw. pornographische oder gewalttätige Abbildungen) aus einer großen Datenmenge herauszufiltern bzw. als solche kenntlich zu machen.

Um zu verstehen, auf welche Art und Weise Accuracy eine Einschätzung der Fähigkeiten des Algorithmus erlaubt, ist es zunächst notwendig zu betrachten, welche Ergebnisse überhaupt ausgegeben werden könnten. Eine Möglichkeit besteht darin, dass der Algorithmus den ihm zur Verfügung gestellten Input, hier also verschiedene Bildmaterialien, korrekt als einer der beiden Kategorien (pornographisch vs. nicht pornographisch) identifiziert. Diese Ergebnisse

KI-Outputs
ACCURACY

	Als Straftat klassifiziert	Nicht als Straftat klassifiziert
Bild beinhaltet Straftat	9	0
Bild beinhaltet keine Straftat	1	0

Korrekt 90% → Accuracy

Inkorrekt 10%

Abb. 4.4 Metadaten Accuracy

werden als *true positive* bezeichnet, wenn ein Bild richtigerweise als pornographischer Inhalt identifiziert wird und *true negative*, wenn ein Bild korrekt als nicht-pornographischer Inhalt klassifiziert wird.

Umgekehrt ist es ebenfalls möglich, dass der Algorithmus den ihm gegebenen Input falsch bewertet: Also etwas als pornographischen Inhalt klassifiziert, das keiner ist, was als *false positive* bezeichnet wird oder einen pornographischen Inhalt übersieht und als „nicht pornographisch" klassifiziert, obwohl es sich um einen handelt, was *false negative* genannt wird.

Betrachten wir nun einen imaginären Algorithmus, dem zehn Inputs gegeben werden: Bilder, die als pornographisch oder nicht-pornographisch klassifiziert werden sollen. Neun der zehn bereitgestellten Inputs sind dabei harmlose Alltagsbilder, nur der zehnte Input zeigt ein Bild mit pornographischem Inhalt. Würde der Algorithmus jetzt alle zehn Bilder als nicht-pornographisch klassifizieren, sähe das Ergebnis wie folgt aus: Neun *true positive* + ein *false positive*, die Accuracy des Algorithmus läge somit bei 90 % (9/10 sind richtig identifiziert worden).

Während 90 % auf den ersten Blick wie ein sehr gutes Ergebnis wirkt, muss der Kontext der Entscheidung und deren Konsequenzen mit in die Beurteilung einbezogen werden, um ein wirkliches Urteil fällen zu können. Was, wenn es sich bei dem als *false positive* gekennzeichneten Input nicht nur um einen störenden pornographischen Inhalt handelt, sondern um eine Person, die im Zuge

der Strafverfolgung nun einem Verbrechen zugeordnet wird, das sie nicht begangen hat oder aber um einen infektiösen Patienten, der als unbedenklich eingestuft wird, obwohl er eine ansteckende Erkrankung hat?

Diese Beispiele verdeutlichen eindringlich, dass Accuracy als Maßeinheit allein nicht ausreichend ist, um wirklich ein Urteil über die Qualität eines Algorithmus zu treffen. Wir benötigen sowohl weitere Metriken als auch ein Verständnis für den Kontext, in dem unsere Ergebnisse für weitere Entscheidungen genutzt werden.

Es sind zusätzliche Metriken notwendig, mit denen die Outputs eines Machine-Learning-Algorithmus interpretiert und bewertet werden können. Zwei, die an dieser Stelle kurz vorgestellt werden sollen sind *Precision* und *Recall* (siehe Abb. 4.5), die beide in direkter Verbindung mit der zuvor präsentierten Accuracy stehen.

Precision beschreibt dabei, wie gut der Algorithmus darin ist, richtige Positive unter allen Positiven zu erkennen. Die Formel lautet:

True Positive/(True Positive + False Positive). Vereinfacht: True Positive/All Positive.

Für das vorherige Beispiel würde dies ebenfalls bedeuten, 9 Inputs sind als *true positives* erkannt worden, also Bilder, die korrekt als harmlose Alltagsbilder

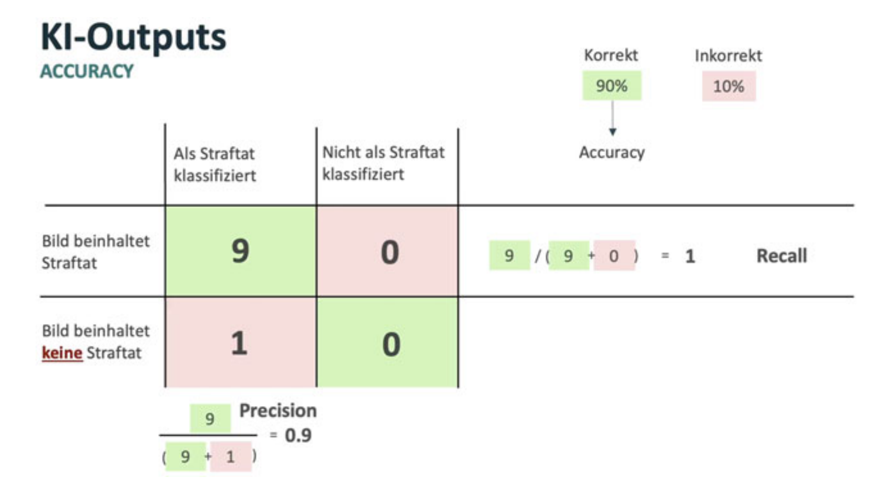


Abb. 4.5 Metadaten Precision & Recall

klassifiziert worden sind, während 1 Input als *false positive*, also als harmloses Bild klassifiziert worden ist, obwohl es pornographische Inhalte abbildet.

Der Formel folgend bedeutet dies eine Precision von 90 %. Precision liefert also einen Wert, der es ermöglicht, eine realistischere Einschätzung darüber vorzunehmen, wie hoch die Wahrscheinlichkeit eines *false positive* ist. Dies muss dann abhängig des jeweiligen Kontextes beurteilt werden. Wie zuvor bereits beschrieben, kann ein *false positive* bei der Sortierung von Tierbildern als zu vernachlässigen akzeptiert werden, während es bei der Überprüfung von potenziellen Straftätern von großer Bedeutung ist.

Recall folgt dabei derselben Logik wie Precision. Hier lautet die Formel allerdings:

True Positive/(True Positive + False Negative) also, wie hoch die wirkliche Anzahl positiver Ereignisse war, die durch den Algorithmus bewertet worden sind.

Recall beschreibt also, wie viele tatsächliche *true positive* vom Algorithmus auch als solche erkannt worden sind. Es ist somit die Metrik, die genutzt werden sollte, wenn ein *false negative* Ergebnis mit hohen Kosten oder Schäden verbunden ist. Dies ist bspw. bei der Erkennung von Krankheiten oder der Einschätzung von Betrugsversuchen gegeben. Hier sind die Konsequenzen einer fehlerhaften Zuordnung so hoch, dass es wichtig ist, eine Einschätzung der Fähigkeiten des Algorithmus zu bekommen, bevor dieser praktisch eingesetzt werden kann.

Die letzte Metrik, der sich im Rahmen dieses Kapitels zugewandt werden soll, ist der sogenannte F1-Score (F1-Wert). Dieser wird genutzt, wenn eine Balance zwischen Precision und Recall hergestellt werden soll. Also, wenn sowohl *false positives*, aber auch *false negatives* im Kontext der Beurteilung relevant sind. Dies kann bspw. der Fall sein, wenn eine besonders ungleiche Verteilung der Outputs vorliegt. Während Accuracy unabhängig davon genutzt werden kann, ob ein Modell in den Bereichen Recall und Precision schlecht performed, leidet der F1-Wert darunter, wenn ein Ungleichgewicht vorliegt. Somit ist Accuracy immer dann nicht isoliert zu betrachten, wenn die assoziierten Risiken von *false negative* und *false positive* unterschiedliche Effekte haben.

Übung
In diesem Abschnitt wurden das Thema Metadaten behandelt. Die folgenden Aufgaben sollten bearbeitet werden bevor man fortfährt:

1. Vergleichen Sie, welche Rolle Accuracy bei den folgenden Systemen spielt und überlegen Sie, auf welchen der entsprechenden Werte würden Sie besonders achten?
   a) Ein System zur Einteilung von Mails in Spam oder Nicht-Spam
   b) Ein System zur Dunkelverarbeitung von Widersprüchen bei Tempolimitverstößen
   c) Ein System zur Berechnung der Statik bei Neubauten
   d) Memoriali zur Genehmigung von Baumaßnahmen am Altbau
2. Sie unterhalten sich mit einer Kollegin über ein System, dass bei sehr komplexen Finanzkonstruktionen hilft, den Überblick zu bewahren und Vorschläge dazu macht, welche Prüfungen als nächstes vorgenommen werden sollten. Sie merkt an, dass das System Schlagzeilen gemacht hat, weil es nur eine 75 % Accuracy besitzt und sie es daher nicht einsetzt. Wie kommentieren Sie das?
3. Welche der folgenden Aussagen ist falsch
   a) Die Accuracy im Bereich Unsupervised Learning sollte immer größer sein als 70 %.
   b) Die Accuracy liegt immer bei mindestens 50 %, da es die Ratewahrscheinlichkeit ist.
   c) Auch Systeme mit Accuracy von besser als 90 % können ungeeignet für bestimmte Aufgaben sein.
   d) Die Accuracy ist ein genauerer Wert als der F1-Score.

4.9 Abschluss

In diesem Kapitel wurde gezeigt, dass die Ergebnisse von Machine Learning und KI-Anwendungen sehr viele verschiedene Formen annehmen können. Die Art und Weise des Ergebnisses beeinflusst zentral, wie ein intelligentes System eingesetzt werden kann. Bevor ein solches System eingesetzt wird, ist es wichtig zu verstehen, welches Ergebnis von dem System erwartet wird und wie damit weitergearbeitet wird. Das erlaubt auch, zu überlegen, wie sich Arbeitsprozesse verändern.

Als erstes wurden dabei Kategorien vorgestellt – das sind gleichwertige Gruppen, zu denen ein Ergebnis einsortiert werden kann, z. B. „Angenommen“ oder „Abgelehnt“.

Danach wurde das Pattern Matching vorgestellt. Dabei wird der Verlauf von Daten betrachtet und Mustern zugeordnet, z. B. wie sich Nutzende in einer Anwendung orientieren.

Im Rahmen von Numerischer Prädiktion wurden Ergebnisse vorgestellt, die sich anders als Kategorien in einem Zahlenraum bewegen und so z. B. andere Lerntechniken zulassen. Ein Beispiel dafür ist die Berechnung der Steuerschuld.

Ein großer Bereich ist auch das Erzeugen synthetischer Ergebnisse. Dabei werden Regeln aus bestehenden Produkten, z. B. Bescheiden, extrahiert, um auf Basis weniger Eingangsinformationen neue Ergebnisse erzeugen zu können.

Das Forecasting von z. B. der Entwicklung von Finanzdaten oder E-Mail-Aufkommen anhand einer Zeitserie wurde ebenfalls behandelt.

Abschließend wurden Accuracy, Precision und Recall als Werte zur Beurteilung der Genauigkeit von Ergebnissen besprochen.

Nutzen Sie die Inhalte des Kapitels, um in den weiteren Kapiteln zu überlegen, welche Auswirkung bestimmte Ergebnisse z. B. auf die Nachvollziehbarkeit, die Automatisierbarkeit oder gar die rechtliche Grundlage eines Systems haben (z. B. wenn es um Urheberrechte geht).

Literatur

Carrasco Kind, M., & Brunner, R. J. (2014). SOM z: Photometric redshift PDFs with self-organizing maps and random atlas. *Monthly Notices of the Royal Astronomical Society, 438*(4), 3409–3421.

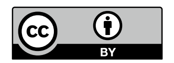
Russell, S., & Norvig, P. (2012). *Künstliche Intelligenz* (Bd. 2). Pearson Studium.

Sokolova, M., Japkowicz, N., & Szpakowicz, S. (2006). Beyond accuracy, F-Score and ROC: A family of discriminant measures for performance evaluation. In A. Sattar & B. Kang (Hrsg.), *AI 2006: Advances in artificial intelligence. AI 2006. Lecture notes in computer science* (Bd. 4304). Springer. https://doi.org/10.1007/11941439_114.

Weiterführende Literatur

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Commun. ACM, 60*(6), 84–90. https://doi.org/10.1145/3065386.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. *arXiv preprint* arXiv:2204.06125.


BY



KI-Strategie 5

Zusammenfassung

Dieses Kapitel adressiert KI-Strategien als Planungs- und Führungsinstrumente (5.1) und gibt Einblick in die Handlungsfelder auf Management-Ebene (5.2).

5.1 KI-Strategien als Planungs- und Führungsinstrumente

Um die Digitale Transformation im öffentlichen Sektor voranzubringen, werden häufig Strategiepapiere erstellt. Wenn nun KI-basierte Systeme vermehrt eingesetzt werden, stellt sich die Frage, inwiefern es einer Strategie bedarf. Was genau ist darunter zu verstehen? Leitbilder, Masterpläne, Roadmaps usw. sind Planungs- und Führungsinstrumente. Sie zeigen, dass es einen ganzheitlichen Blick auf die Thematik allgemein, die Ziele, die Rahmenbedingungen, die Herausforderungen und auch die Lösungsansätze gibt. Sie vermitteln außerdem ein Bild von der Zukunft und dienen als Orientierung in der Umsetzung – im Fall von Masterplänen und Roadmaps mit konkreten Umsetzungsplänen. Strategiepapiere dienen auch der Kommunikation und letztlich der Profilierung. Sie zeigen, dass sich Organisationen mit bestimmten Themen intensiv befassen, schon bevor konkrete Ergebnisse sichtbar sind. Bei der Betrachtung von Strategiepapieren können wir uns an den Ebenen des St. Galler Managementmodells orientieren (vgl. zu folgenden Ausführungen Bleicher, 2011, S. 87 ff.). Die normative Ebene stellt die handlungsbegründende Ebene dar. Hier werden Missionen vorgegebenen, die der Organisationsentwicklung dienen. Die strategische Ebene fokussiert die grundlegende Ausrichtung der Organisation bezogen auf Strukturen, Leistungen, den

M. Heine et al., *Künstliche Intelligenz in öffentlichen Verwaltungen*, Edition eGov-Campus, https://doi.org/10.1007/978-3-658-40101-6_5

Umgang mit Ressourcen und Aktivitäten. Aus der Mission wird ein Programm. Und schließlich adressiert die operative Ebene den konzeptgeleiteten Vollzug. Zusätzlich gibt es die Vision als Sonderfall. Sie durchdringt die drei Ebenen und fungiert als Impulsgeber und Leitstern. Je nach Ebene können Strategiepapiere als Steuerungsinstrumente sehr unterschiedliche Wirkungen entfalten. Es kann um einen eher allgemeinen Rahmen und abstrakte Zielsetzungen gehen oder um sehr konkrete Handlungsschritte mit einer festen Zeitschiene. Betrachten wir nun Steuerungsinstrumente für KI im öffentlichen Sektor. Grundlegend ist, dass diese Instrumente nicht isoliert stehen, sondern eng an die Mission der Organisation gebunden sind und genau diesen Zielen dienen. Mit Blick auf die Nationale E-Government-Strategie in Deutschland ist das die konsequente Ausrichtung an den Staatszielen und dem geltenden Recht. Eine KI-Strategie muss sich also zum einen ebenso konsequent an der Mission der Organisation orientieren und zum anderen müssen auch Beziehungen zu anderen Strategien, zum Beispiel E-Government-Strategien, Berücksichtigung finden (siehe für Beispiele Abb. 5.1). Insbesondere müssen mögliche Zielkonflikte identifiziert und abgewogen werden.

Eine zentrale Frage auf abstrakter Ebene ist also: Inwiefern dient die Anwendung von KI-basierten Systemen den Staatszielen. Eine Orientierung am Gemeinwohl spielt hier eine zentrale Rolle. Darunter fällt insbesondere auch

Ziele der Nationalen E-Government-Strategie 2015	Konkretisierung	Beispiele für KI-bezogene Handlungsfelder
Nutzen für Bürger, Unternehmen und Verwaltung	Zugang, E-Government-Kompetenzen	Beratung 24/7 durch Chatbots und Sprachassistenten, automatisierte Übersetzung
Wirtschaftlichkeit, Effizienz und Leistungsfähigkeit	Prozessmanagement, Zusammenarbeit, Modulare IT-Architekturen	Automatisierung von Teilaufgaben, KI-basiertes Prozessmanagement, KI-Architekturen
Informationssicherheit und Datenschutz	Schutzmaßnahmen, Resilienz (Funktionsfähigkeit in Krisenzeiten)	Risikomanagement, Informationspflichten und Auskunftsrechte, Folgeabschätzung
Transparenz und gesellschaftliche Teilhabe	Open Data und Informationsfreiheit, Partizipation	KI-Anwendungen zur Erhöhung der Teilhabe (z. B. in der Analyse von Beiträgen in Partizipationsverfahren)
Zukunftsfähigkeit und Nachhaltigkeit	Innovationsfähigkeit, Veränderungsbereitschaft, Wiederverwendbarkeit, ökologische Nachhaltigkeit	KI-Weiterbildung, Innovations- und Entwicklungsprogramme zu KI-Anwendungen

Abb. 5.1 Ziele der Nationalen E-Government Strategie aus dem Jahr 2015 und KI-bezogene Handlungsfelder

operativer Nutzen, also effektive und effiziente Prozesse. Die Ziele sind klar zu definieren. Wollen wir Prozesse effizient gestalten? Wollen wir eine erhöhte Automatisierung von Prozessen, um trotz knapper Personalressourcen leistungsfähig sein zu können? Wollen wir Beschäftigte entlasten – von belastenden Tätigkeiten? Und mit Blick auf die Servicequalität bei Verwaltungsleistungen: Ist es Ziel, eine bessere Erreichbarkeit oder stets eine freundliche und individuelle Ansprache zu gewährleisten? Darüber müssen wir uns im ersten Schritt klarwerden, um überhaupt eine tragfähige KI-Strategie entwerfen zu können, Handlungsschwerpunkte zu setzen und später die Erreichung der Ziele überprüfen zu können.

5.2 KI-Management

Der Einsatz von KI im öffentlichen Sektor geht mit einer Vielzahl von Herausforderungen einher, die vielfältige Organisationsbereiche und -funktionen betreffen und durch Management-Fragen adressiert werden können (van Giffen et al., 2020):

- Welche Implikationen hat KI innerhalb unserer Strategie? Wie können KI-Projekte sicher und effizient durchgeführt werden? Welche Voraussetzungen müssen geschaffen werden?
- Welche Fähigkeiten und welche Organisationsstruktur brauchen wir, um die Potenziale von KI voll auszuschöpfen?
- Welche rechtlichen, regulatorischen und vertraglichen Grundlagen müssen wir in der produktiven Nutzung von KI berücksichtigen?
- Wie könnte ein KI-Lebenszyklus-Management aussehen, das sich gut in unsere bestehenden Prozesse einfügt?
- Wie müssen Prozesse für Auswahl, Implementierung und Betrieb einer KI-Technologie-Infrastruktur definiert sein?
- Welche Maßnahmen sind geeignet, um Cybersicherheit herzustellen und gegnerische Angriffe auf KI-Systeme effizient abzuwehren?

Die Beantwortung dieser Fragen kann begleitend zu ersten Projekten erfolgen. Auf diese Weise wird bereits Expertise gewonnen und kann in die Beantwortung – und somit in die KI-Governance, also die ganzheitliche Steuerung – einfließen.

Basierend auf dem St. Galler Management Modell sind folgende Handlungsfelder bei der Steuerung des KI-Einsatzes relevant (van Giffen et al., 2020):

1. Management von Künstlicher Intelligenz
2. Organisation des Betriebs
3. Rechtliche Gestaltung
4. Regulierung und Compliance
5. Lebenszyklus-Management
6. Management der Technologie-Infrastruktur
7. Cybersicherheit

Mit Blick auf die Anwendung in öffentlichen Verwaltungen befasst sich das Management von KI (1.) mit den notwendigen Organisationsstrukturen und Management-Prozessen, um die definierten KI-bezogenen Ziele der Behörde umzusetzen. Erste Projekterfahrungen bilden die Basis für die systematische Gestaltung des Management-Modells. Diese Erfahrungen beziehen sich zum einen auf das KI-System selbst, zum anderen jedoch auch auf Voraussetzungen (zum Beispiel notwendige IT-Infrastruktur, KI-Kompetenzen, Verfügbarkeit relevanter Daten), Grenzen, aber auch Erkenntnisse zum Projektmanagement. Auf Basis erster (und weiterer) Projekterfahrungen sind Beschäftigte außerdem in der Lage, KI-Anwendungsoptionen für das eigene Arbeitsumfeld zu identifizieren und zu reflektieren. Zentral ist also die Durchführung von KI-Projekten wie auch die Analyse der Projekterfahrungen, um Organisationales Lernen zu ermöglichen.

Das Handlungsfeld Organisation des Betriebs (2.) adressiert die Fähigkeit, geeignete KI-Anwendungsfälle zu identifizieren. Außerdem muss sichergestellt werden, dass die nötigen KI-Kompetenzen in der Organisation aufgebaut werden. Zu berücksichtigten ist auch die Rolle externer Dienstleister und Technologieanbieter. Es ist zu klären, welche Dienste extern bezogen werden sollen, wie die Zusammenarbeit mit dem Dienstleister zu gestalten ist und welche Anforderungen dabei relevant sind. Diese Frage ist auch vor dem Hintergrund der Digitalen Souveränität der öffentlichen Verwaltung relevant und wird im nächsten Lern-Item behandelt.

Die rechtlichen Gestaltungsfragen (3.) umfassen im öffentlichen Sektor grundsätzliche regulatorische Anforderungen (vgl. Lerneinheit KI und Recht in diesem Kurs) und Maßnahmen zur Sicherstellung ihrer Berücksichtigung (4. Compliance).

Im Handlungsfeld Lebenszyklus-Management (5.) steht die Frage, wie das KI-System wirkungsvoll in die Leistungserstellung integriert werden kann. Im Fokus

steht die Entwicklung von KI-Systemen, vom Prototyp hin zum Produktivsystem. Welches Problem löst das System? Welchen Funktionsumfang soll das System haben? Welche Daten bilden die Grundlage? Welche KI-Methoden kommen zum Einsatz? Wie wird das System evaluiert und weiterentwickelt? Die Vielfalt der Aufgaben zeigt, dass spezifische Rollen zu etablieren sind. Das sind zum Beispiel Data Scientists, KI-Spezialisten, Fachvertreter und Projektmanager.

Grundlegend für KI-Projekte ist der Aufbau einer entsprechenden Infrastruktur (6.). Dazu gehört die Auswahl geeigneter Frameworks und Bibliotheken, aber auch geeigneter Hardware. Die Auswahl sollte auf einer systematischen Anforderungsanalyse beruhen, wobei auch hier in den ersten KI-Projekten wertvolle Erfahrungen aufgebaut werden können.

Im Zuge der IT-Sicherheit (7.) sind ebenfalls einige Fragen zu beantworten (BSI, 2021): Welche Angriffsszenarien entstehen durch den Einsatz von KI und welche Auswirkungen haben diese? Wie können KI-Systeme gegen KI-spezifische Angriffe nachweisbar geschützt werden? Darüber hinaus bietet der Einsatz von KI auch Möglichkeiten, IT-Sicherheit zu verbessern. Auf der anderen Seite entstehen neue Bedrohungen für IT-Systeme, da KI-Methoden auch in Angriffswerkzeugen genutzt werden können.

Insgesamt zeigt sich, dass mit dem Einsatz von KI-Systemen Herausforderungen verbunden sind, die auf Management-Ebene systematisch adressiert werden müssen und im Zuge der zunehmenden KI-Projekterfahrungen stetig anzupassen sind.

5.3 Aufgaben zum eigenen Anwendungsfalls

Nach dem Grundlagenteil ist es nun Zeit, noch einmal auf Ihr gewähltes KI-Projekt zu schauen. Sie kennen nun auch die Funktionen von Strategiepapieren, Ziele des KI-Einsatzes und Handlungsfelder auf Management-Ebene. Lassen Sie dieses Wissen in die Auswahl der KI-Anwendungsbereiche und des konkreten KI-Projektes einfließen und bearbeiten Sie folgende Teilaufgaben:

- Beschreiben Sie für die Organisation eine KI-Vision. Diese Aufgabe bezieht sich nicht auf das konkrete KI-Projekt, sondern auf die Organisation insgesamt.
- Wählen Sie einen geeigneten Anwendungsbereich für den KI-Einsatz in Ihrer Organisation (bzw. überdenken Sie Ihre getroffene Entscheidung und passen Sie sie ggf. an).
- Begründen Sie die Auswahl des Anwendungsbereichs.

- Beschreiben Sie außerdem Ziele auf normativer und strategischer Ebene. Diese Aufgabe bezieht sich nicht auf das konkrete KI-Projekt, sondern auf die Organisation insgesamt.
- Überdenken Sie nun Ihr gewähltes KI-Projekt und nehmen Sie ggf. Anpassungen vor.
- Detaillieren Sie auch die Ausführungen in Bezug auf die Kap. 2 bis 4.

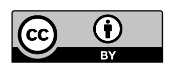
Literatur

Bleicher, K. (2011). *Das Konzept Integriertes Management: Visionen – Missionen – Programme*. Campus.

BSI (2021). *Künstliche Intelligenz*. Bundesamt für Sicherheit in der Informationstechnik. https://www.bsi.bund.de/DE/Themen/Unternehmen-und-Organisationen/Informationen-und-Empfehlungen/Kuenstliche-Intelligenz/KI.html?nn=129146. Zugriffen: 23. Dez. 2021.

van Giffen, B., Borth, D., & Brenner, W. (2020). Management von Künstlicher Intelligenz in Unternehmen. *HMD Praxis der Wirtschaftsinformatik, 57*(1), 4–20. https://doi.org/10.1365/s40702-020-00584-0.

Nationale E-Government-Strategie. (2015). IT-Planungsrat. https://www.it-planungsrat.de/der-it-planungsrat/nationale-e-government-strategie. Zugegriffen: 31. Aug. 2022.



Gebrauchstaugliche Entwicklung von KI-Anwendungen

6

Zusammenfassung

Dieses Kapitel behandelt die Gebrauchstauglichkeit von Produkten, d. h. welche Eigenschaften ein Produkt wie eine KI-Anwendung aufweisen muss, damit man damit gut arbeiten kann. Der menschzentrierte Gestaltungsprozess wird als ein Weg dargestellt, solche gebrauchstauglichen Anwendungen zu entwickeln. Dabei wird auf die Rolle der zukünftigen Nutzer, die besonderen Anforderungen von KI-Anwendungen und der öffentlichen Verwaltung eingegangen.

6.1 Einleitung

Wie muss eine KI-Anwendung – oder jegliches Produkt – gestaltet sein, dass man „gut" damit arbeiten kann? Und woran kann man „gut" festmachen? Um diese Fragen zu beantworten werden zunächst drei Fallbeispiele vorgestellt (Abschn. 6.2) und der Begriff der Gebrauchstauglichkeit (Abschn. 6.3) definiert – was unter „gut arbeiten" zu verstehen ist. Um gebrauchstaugliche Anwendungen zu entwickeln, bietet sich der menschzentrierte Gestaltungsprozess an, der die zukünftigen Nutzer und andere Betroffene frühzeitig einbezieht (Abschn. 6.4). Dabei stellen sowohl KI-Anwendungen (Abschn. 6.5) als auch die öffentliche Verwaltung (Abschn. 6.6) besondere Anforderungen. Um Ihnen die Möglichkeit zu geben, KI-Anwendungen zu bewerten, werden mögliche Frage vorgestellt, die Sie an KI-Anwendungen in der ÖV stellen können (Abschn. 6.7), gefolgt von einer Bitte, sich bei der menschzentrierten Entwicklung zu beteiligen (Abschn. 6.8). Zur Lernüberprüfung folgen Übungsfragen (Abschn. 6.9) und Aufgaben zum eigenen Anwendungsfall (Abschn. 6.10).

M. Heine et al., *Künstliche Intelligenz in öffentlichen Verwaltungen*, Edition eGov-Campus, https://doi.org/10.1007/978-3-658-40101-6_6

6.2 Fallbeispiele

Betrachtet man KI-Anwendungen daraufhin, inwieweit man gut mit ihnen arbeiten kann, lassen sich gute, schlechte und hässliche Anwendungen identifizieren.

Gute Anwendungen wie z. B. „Spotify“ (Streamingdienst für Musik, der mittels KI passende Musik für die Nutzer auswählt) „funktionieren einfach“. Sie hören Musik und bemerken nicht unbedingt, dass im Hintergrund ein System Ihre Musikpräferenzen lernt. Die Anwendung spielt einfach gute Musik. Sie erreichen auf einfache Art Ihr Ziel, gute Musik zu hören.

Schlechte Anwendungen wie z. B. Microsoft’s Clippy („Karl Klammer“, eine digitale animierte Büroklammer in einer frühen Microsoft Word-Version) versuchen zu helfen, erreichen aber das Gegenteil. Clippy erkannte zum Beispiel, wenn der Nutzer einen Brief schreibt – die Versuche, das Schreiben des Briefes zu unterstützen, waren jedoch eher hinderlich. Da das System erst erkennen konnte, was geschrieben wurde, wenn der Nutzer schon mit dem Schreiben angefangen hat, hat die Animation zu Beginn des Schreibprozesses Aufmerksamkeit auf sich gezogen und den Prozess unterbrochen. Anstatt zu unterstützen, das Ziel zu erreichen, hat das System den Nutzer behindert und frustriert.

Hässliche Anwendungen wie der fiktionale Computer „HAL 9000“ aus „2001: A Space Odyssey“ behindern den Nutzer nicht nur, sie arbeiten aktiv gegen den Nutzer und nehmen ihm Autonomie und Kontrolle. Auch wenn es sich hierbei um ein Science-Fiction-Beispiel handelt, gibt es – weniger gravierend – schon heute eingesetzte Anwendungen, welche die Möglichkeiten der Nutzer einschränken. Im einfachen Fall wird einem Sachbearbeiter zum Beispiel der Ermessensspielraum genommen, weil dies in der Anwendung nicht vorgesehen ist. Mit entsprechend „hässlichen“ Konsequenzen für den Antragsteller.

Die Aufgabe einer gebrauchstauglichen Entwicklung von KI-Anwendungen ist es, Anwendungen zu entwickeln, die „einfach funktionieren“, und Anwendungen wie Clippy oder gar HAL 9000 zu vermeiden.

6.3 Gebrauchstauglichkeit

Wenn gesagt wird, dass man mit einer Software – oder generell mit einem Produkt – „gut arbeiten“ kann, was genau ist damit gemeint? Um diese Anforderungen genauer zu definieren, bietet sich der Begriff der „Gebrauchstauglichkeit“ an. Dieser ist in der Norm EN ISO 9241-210:2010 wie folgt definiert:

►„**Gebrauchstauglichkeit**: Ausmaß, in dem ein System, ein Produkt oder eine Dienstleistung durch bestimmte Benutzer in einem bestimmten Nutzungskontext genutzt werden kann, um festgelegte Ziele effektiv, effizient und zufriedenstellend zu erreichen.“ EN ISO 9241-210:2010.
Neben den drei Kriterien effektiv, effizient und zufriedenstellend ist zusätzlich – nach einer Definition von Batinic und Appel (2008) – noch die Erlernbarkeit wichtig. Damit ergeben sich folgende Kriterien für „gut arbeiten“:

- **Effektivität (Zielerreichung):** Können die Ziele erreicht werden, z. B. einen Antrag zu bearbeiten?
- **Effizienz (Aufwand; schnell & ohne Fehler):** Können die Ziele mit wenig Aufwand, d. h. schnell und ohne Korrekturen erreicht werden?
- **Erlernbarkeit (einfach erlernbar):** Ist die Arbeit mit der Anwendung leicht erlernbar – und findet man sich schnell wieder zurecht, wenn man sie lange nicht bedient hat?
- **Zufriedenstellung (positive Einstellung):** Hat man ein zufriedenstellendes Gefühl, wenn man mit der Anwendung arbeitet?

Gebrauchstauglichkeit zu gewährleisten, ist alles andere als einfach. Selbst bei Benutzeroberflächen (Interfaces), die „offensichtlich“ zu bedienen sind, kann es zu Fehlern bei der Benutzung kommen. So kann z. B. der Fingerabdrucksensor eines iPhones trotz einer augenscheinlich eindeutigen Abbildung missverstanden werden und der Finger auf den Screen, statt auf dem Home-Button, gelegt werden. Ein kleiner Fehler, den die Person zwar selbst beheben kann, aber ein gutes Beispiel, dass es selbst bei Weltmarktführern wie Apple zu Problemen in der Gebrauchstauglichkeit kommen kann.

Ein anschauliches Beispiel für eine fiktive Anwendung, bei der alle vier vorgestellten Gebrauchstauglichkeitskriterien verletzt sind, sieht man in dem Kurzfilm „Lifted“ von Pixar. Sie finden das Video z. B. auf YouTube. In diesem Kurzfilm versucht ein Außerirdischer einen Menschen zu entführen – scheitert dabei aber an einer nicht gebrauchstauglichen Benutzeroberfläche. Die vielen kleinen Schalter machen die Anwendung schwer zu erlernen (schlechte Erlernbarkeit). Sie führen – zum Leidwesen des Menschen – zu einem langwierigen Prozess mit vielen Fehlern (nicht effizient). Das führt letztlich dazu, dass das Ziel nicht erreicht wird (nicht effektiv). Und insgesamt tritt dadurch mangelnde Zufriedenheit auf (keine Zufriedenstellung). Ein fiktives Beispiel mit einer Benutzeroberfläche, die so nie konstruiert werden würde. Es bringt durch die Überspitzung allerdings die Frustration vieler Nutzer mit schlecht entwickelten, d. h. nicht gebrauchstauglichen, Anwendungen auf den Punkt.

Um zu beurteilen, inwieweit eine Anwendung gebrauchstauglich ist, müssen die Kriterien genau erfasst werden. Es gibt verschiedene Möglichkeiten, die Kriterien der Gebrauchstauglichkeit zu messen (zu „operationalisieren", d. h. Operationen angeben, wie diese erfasst werden können). Dazu gehören z. B. Beobachtungen (messen ob die Ziele erreicht wurden und wieviele Fehler auftreten), Eye Tracking (messen wohin die Person auf dem Bildschirm schaut, z. B. ob die relevanten Informationen gesehen wurden), oder Fragebögen (Bewerten der Gebrauchstauglichkeit über Selbstauskünfte). Insbesondere Verhaltensmaße (im Gegensatz zu Selbstauskünften bei Fragebögen) sind hilfreich um zu überprüfen, z. B. ob eine Person eine Aufgabe wirklich erfolgreich bearbeitet und wieviel Zeit sie dafür benötigt hat. Fragebögen haben allerdings den Vorteil, dass sie sehr gut skalieren – man kann schnell hunderte von Nutzern befragen. Ein älterer aber häufig eingesetzter Fragebogen ist die System Usability Scale (SUS). Er besteht aus zehn Fragen (z. B. ob das System einfach zu benutzen ist, schnell zu erlernen ist, etc.). Der Vorteil dieses Fragebogens ist, dass die Auswertung klar vorgegeben ist und ein Bewertungsmaßstab vorgegeben wird (vgl. Bangor et al., 2008).

Wie kann man gewährleisten, oder zumindest wahrscheinlicher machen, dass eine Anwendung auch gebrauchstauglich ist? Dafür bietet sich der menschzentrierte Gestaltungsprozess an.

6.4 Menschzentrierte Gestaltung

Wie kann man Anwendungen entwickeln, die auch wirklich gebrauchstauglich sind? Ein Weg ist die Verwendung des menschzentrierten Gestaltungsprozesses nach der Norm „Ergonomie der Mensch-System-Interaktion Teil 210: Prozess zur Gestaltung gebrauchstauglicher interaktiver Systeme" (DIN EN ISO 9241-210, 2011).

Der Prozess beginnt mit dem Planen des menschzentrierten Gestaltungsprozesses (z. B. Methoden, Zeitplan und Ressourcen festlegen). Danach versucht man, den Nutzungskontext zu verstehen und legt ihn genauer fest (wer sind die Benutzer, welche Aufgaben sollen getätigt werden, etc.). Daraus ergeben sich die Nutzungsanforderungen, die im folgenden Prozess immer wieder zur Bewertung der Qualität der Lösungen verwendet werden. Sie sind auch die Basis für den nächsten Schritt, die Erarbeitung von Gestaltungslösungen, welche die Nutzungsanforderungen erfüllen sollen. Dazu gehören z. B. die Erarbeitung von Zeichnungen bis hin zu bedienbaren Anwendungen. Diese Lösungen werden auf

Basis der Nutzungsanforderungen bewertet (evaluiert). Dies geschieht schon während des Gestaltungsprozesses, indem z. B. erste Skizzen der Anwendungen von der Zielgruppe bewertet werden und das Feedback in die Überarbeitung eingeht (formative Evaluation).

Kernbestandteil dieses Prozesses sind die Iterationen – man erwartet nicht, dass der erste Entwurf einer Anwendung direkt die Nutzungsanforderungen gut erfüllt. Stattdessen ist eingeplant, und erwünscht, zu früheren Schritten zurückzugehen. Dazu gehört das Verstehen und Festlegen des Nutzungskontextes, wenn sich in der Evaluation herausstellt, dass z. B. bestimmte Personengruppen oder Bedingungen nicht ausreichend bekannt sind (z. B. wenn sich bei einer Gestaltungslösung herausstellt, dass Prozesse in der Praxis anders ausgeführt werden). Es können auch die Nutzungsanforderungen überarbeitet werden (z. B. Anforderungen hinzufügen oder anders priorisieren) oder neue Gestaltungslösungen auf Basis des Evaluationsfeedbacks erarbeitet werden. Über meist mehrere Iterationen wird so eine Anwendung entwickelt, welche die Nutzungsanforderungen erfüllt. Diese Anwendung kann dann in einer abschließenden (summativen) Evaluation von einer größeren Gruppe von Nutzern bewertet werden.

Zentral im menschzentrierten Gestaltungsprozess ist, dass bei all diesen Schritten die Nutzer involviert werden – sowie die Personen, die von der Anwendung betroffen sind (Stakeholder). So werden bei einem Entscheidungsunterstützungssystem in der Sozialverwaltung (ein KI-System schlägt beispielsweise vor, welchen Anträgen stattgegeben werden sollte) z. B. die Sachbearbeiter zum Verstehen und Festlegen des Nutzungskontextes, der Bestimmung der Anforderungen und Bewertung der Lösungen einbezogen. Werden Bescheide z. B. automatisch verschickt, sollten als Betroffene auch die Antragsteller mit einbezogen werden (z. B. inwieweit der automatisch generierte Begründungstext für eine Ablehnung nachvollziehbar ist). Es sind die Nutzer, und die Betroffenen, die frühzeitig und wiederholt die Gebrauchstauglichkeit der Anwendung bewerten. Ihr Feedback wird in nachfolgenden Entwürfen berücksichtigt, um die Anwendung immer besser zu gestalten, bis sie schließlich die Anforderungen gut erfüllt.

Praktisch kann ein solcher menschzentrierter Gestaltungsprozess in die Phasen der Analyse, Konzeption (mit formativen Evaluationen), Realisierung und abschließende Evaluation unterteilt werden.

6.4.1 Analysephase

In der Analyse versucht man, den Nutzungskontext und die Anforderungen genau zu verstehen. Dabei helfen unter anderem die folgenden Analysen:

- **Benutzeranalyse:** Wer sind die Personen, welche die Software verwenden bzw. davon beeinflusst werden (Stakeholder)? Was kennzeichnet sie? Was muss berücksichtigt werden?
- **Problem-/Aufgabenanalyse:** Welche Aufgaben sollen mit der Software bearbeitet werden? Welche Probleme gilt es zu überwinden? Wie sieht die Arbeitstätigkeit für den Nutzer aus?
- **Organisationsanalyse:** Welche organisationalen oder rechtlichen Rahmenbedingungen müssen beachtet werden?
- **Kontextanalyse:** Wie ist der räumliche und zeitliche Kontext bei der Nutzung? Was muss hier beachtet werden?

Beispiel: Benutzeranalyse

Bei der Benutzeranalyse sind u. a. Interviews und Umfragen hilfreich, um die Nutzer zu verstehen und angemessen zu berücksichtigen. Bei KI-Systemen sind es v. a. Einstellungen, die über den Erfolg oder Misserfolg einer Software entscheiden können. Aber auch Vorwissen sowie evtl. zu hohe oder falsche Erwartungen müssen adressiert werden. Befragt man Mitarbeiter der öffentlichen Verwaltung zu KI-Systemen, dann kann unter Umständen beobachtet werden, dass viele Mitarbeiter davon ausgehen, dass ein KI-System schnell, objektiv, effizient und um jede Uhrzeit arbeitet. Man kann aber auch feststellen, dass dem KI-System weitgehend der Blick auf den Einzelfall abgesprochen wird. Ein Entscheidungsunterstützungssystem in der Sozialverwaltung, bei denen Mitarbeiter von KI-Systemen unterstützt werden sollen, muss z. B. solche Einstellungen adressieren. Dies kann beispielsweise dadurch passieren, dass ein KI-System nur Vorschläge macht bzw. die Erfüllung von Kriterien beurteilt, der Mitarbeiter aber weithin Ermessensspielraum nutzen kann.◄

Insbesondere bei der Einführung von KI-Systemen in Verwaltungsbereiche kann es hilfreich sein, die Einstellung der Mitarbeiter zu KI zu klassifizieren. Zhu und andere (2021) haben dafür ein Modell entwickelt, welches Personen in

vier Quadranten gruppiert. Es verwendet eine rationale (kognitive, was denken die Personen über KI) und eine emotionale (was ist das „Bauchgefühl“ bei KI) Dimension. Die „KI-Furchtlosen“ sehen die Vorteile und sind optimistisch, was KI betrifft. Die „KI-Skeptiker“ haben ebenfalls positive Emotionen und sind interessiert an KI, möchten aber den Wert klar erkennen können. Die „KI-Zurückhaltenden“ sehen zwar kognitiv die Vorteile, KI fühlt sich für diese Personen allerdings nicht gut an, sie sind misstrauisch. Die „KI-Abweichler“ oder „KI-Dissidenten“ sehen weder die Vorteile, noch fühlt sich KI für diese Gruppe positiv an.

Auch wenn Kognition und Emotion keine unabhängigen Dimensionen sind (was Menschen denken, beeinflusst stark was sie fühlen und umgekehrt), hat das Modell einen hohen praktischen Nutzen. So sollten alle Gruppen in der menschzentrierten Entwicklung angesprochen werden und Feedback geben. Ein weiterer Nutzen liegt im Change Management, da es hilft, die unterschiedlichen Gruppen spezifisch anzusprechen. KI-Zurückhaltende sehen zum Beispiel schon die Vorteile – rationale Argumente, dass der Einsatz von KI zu mehr Effizienz führend wird, sind überflüssig. Das ist dieser Gruppe bekannt – sie hat eher emotionale Bedenken, ein schlechtes Bauchgefühl, wohin z. B. der Einsatz von KI führen könnte. Diese Gruppe muss man auf der emotionalen Ebene ernst nehmen und adressieren. Zhu und andere (2021) empfehlen u. a. die Furchtlosen als Multiplikatoren zu nutzen, die Bedenken der Dissenter zu adressieren, den Skeptikern zuzuhören (u. a. da sie konstruktiv-kritisch Probleme identifizieren können), und aus den Zurückhaltenden Furchtlose zu machen, indem man ihre Emotionen ernst nimmt und adressiert. Insbesondere muss man die emotionale Seite ernst nehmen. Bedenken, u. a. bezüglich Arbeitsplatzsicherheit, Arbeitsinhalt, Einfluss von KI auf die Gesellschaft (menschliche Intelligenz wird abgewertet, Wert von Menschen und Menschenwürde), müssen adressiert werden. Zum Beispiel sollten KI-Systeme vor allem menschliche Aktivitäten ergänzen und Autonomie und Kontrolle des Menschen fördern statt nehmen (siehe dazu auch Kap. 7). Einen großen Einfluss hat auch der erste Kontakt mit einer KI-Anwendung im Arbeitskontext. Die ersten Anwendungen müssen flexibel, verlässlich und einfach zu bedienen sein. Zum Teil müssen aber auch Erwartungen im Vorfeld adressiert werden und evtl. Organisationsstrukturen, soziale Normen und Gesetze verändert werden.

Zhu und andere (2021) haben auch analysiert, mit welchen Variablen die Zuteilung in die Quadranten einhergeht. Hierbei stellte sich heraus, dass z. B.

nicht Alter oder Geschlecht den Ausschlag geben, sondern Aspekte wie „Technologieoptimismus“, „wahrgenommene kognitive Fähigkeiten von KI“, „wahrgenommene operative Fähigkeiten von KI“, „vermutete schädliche Auswirkungen von KI“, und die „Wissensintensität der eigenen Arbeit“.

Beispiel: Aufgaben-/Problemanalyse

Bei der Aufgaben-/Problemanalyse wird der Arbeitsprozess genau untersucht und z. B. auf den möglichen Einsatz von KI überprüft. Dieses Vorgehen zeigt die Arbeit von Houy et al. (2020). Die Autoren haben zuerst den Arbeitsprozess ohne KI erarbeitet. Aufgrund der Zerlegung in die Teilschritte kann dann bei jedem Schritt überprüft werden, ob bzw. inwieweit KI-Systeme den Prozessschritt unterstützen können. Der untersuchte Prozess kann z. B. über Handschriftenerkennung (Optical Character Recognition, OCR), Robotic-Process-Automation (RPA), Vollständigkeitsprüfungen (u. a. via Natural Language Processing/NLP) oder Bescheiderstellung via Natural Language Generation unterstützt werden.◀

6.4.2 Konzeptionsphase

In der Konzeptionsphase werden mögliche Gestaltungslösungen entwickelt. Hierbei versucht man möglichst viele verschiedene Ideen zu generieren (divergent zu denken). Diese werden von den Nutzern bewertet und das Feedback wird für neue Gestaltungslösungen aufgegriffen (formative Evaluation). Die besten Ideen werden weiter ausgearbeitet. Die Kriterien für gebrauchstaugliche Anwendungen (effektiv, effizient, zufriedenstellend, leicht erlernbar) sind hier zentral bei der Bewertung.

Als Anwender wird man meist mit Skizzen von möglichen Benutzeroberflächen konfrontiert, zu denen man Feedback geben soll. Die Skizzen wirken dabei einfach und unfertig, weil sie bewusst vorläufig sind. Würde man direkt programmierte oder echt aussehenden Benutzeroberflächen zeigen, ist das Feedback meist zurückhaltender und geht zu stark auf kleine Details ein (z. B. Position von Buttons, Farbgestaltung). Wichtiger ist erst einmal, grundsätzliches Feedback zu erhalten. Es gibt auch eine Reihe von Heuristiken („Daumenregeln“), die bei der Gestaltung zu berücksichtigen sind. Die Bekanntesten stammen von Nielsen (1994).

6.4.3 Realisierungsphase

In der Realisierungsphase wird die Anwendung programmiert. Hierbei versucht man sich möglichst nahe an die beste Idee aus der Konzeption zu halten, wobei auch während oder nach der Entwicklung die Anwendung evaluiert und gegebenenfalls weiter verbessert wird.

6.4.4 Summative Evaluationsphase

Bei der Entwicklung werden die Nutzer, wie beschrieben, in jeder Phase einbezogen. Ihr Feedback ist entscheidend, um eine gebrauchstaugliche – d. h. effektive, effiziente, zufriedenstellende und leicht erlernbare – Anwendung zu entwickeln.

Während der Entwicklung (v. a. in der Konzeptionsphase) passiert dies innerhalb von formativen Evaluationen, bei denen wenige Nutzer schnell Feedback geben. Nach der Realisierung sollte die finale Lösung abschließend (summativ) evaluiert werden. Hierbei wird Feedback von einer größeren Menge von Nutzern eingeholt, idealerweise durch Verwendung der Anwendung im echten Nutzungskontext mit realistischen (oder echten) Aufgaben.

6.4.5 Fazit zur menschzentrierten Gestaltung

Auch wenn der menschzentrierte Gestaltungsprozess nicht garantieren kann, dass die Anwendung nach der Entwicklung für alle Nutzer gebrauchstauglich ist – das kann bei einer kreativen Tätigkeit kein Prozess – macht er die Entwicklung einer solchen gebrauchstauglichen Anwendung wahrscheinlicher. Die Nutzer werden ernst genommen und in die Entwicklung einbezogen. Mögliche Probleme können bereits in frühen Stadien der Entwicklung identifiziert und – mit wesentlich weniger Aufwand als bei einer fertig realisierten Anwendung – behoben werden.

6.5 Besondere Anforderungen bei KI-Anwendungen

KI-Anwendungen müssen – da sie z. B. beim Entscheiden unterstützen oder sogar (teil-)autonom agieren – zusätzlich zu den erwähnten Kriterien von Gebrauchstauglichkeit (Abschn. 6.3; Effektivität, Effizienz, Zufriedenstellung, und Erlernbarkeit, EN ISO 9241-210:2010 und erweitert mittels Batinic & Appel, 2008) noch weitere Anforderungen erfüllen. Die Anforderungen für KI-Anwendungen

lassen sich unter dem Begriff „Vertrauenswürdigkeit“ zusammenfassen (Poretschkin et al., 2021) und umfassen Fairness, Autonomie und Kontrolle, Transparenz, Verlässlichkeit, Sicherheit, sowie Datenschutz.

Der Prüfkatalog von Poretschkin und anderen (2021) gibt hier eine sehr detaillierte Übersicht dieser Kriterien. Bei der Berücksichtigung der Kriterien müssen sowohl Nutzer und Betroffene, aber auch z. T. Experten berücksichtigt werden (z. B. wenn es um die Transparenz eines Systems geht). Speziell bei Anwendungen, die maschinelles Lernen (ML) verwenden, ist zu beachten, dass damit häufig probabilistische Entscheidungen getroffen werden. Die Qualität, z. B. der Unterstützung, variiert dadurch, was eine Bewertung (Evaluation) der Anwendung erschwert. Auch gibt es ML-Anwendungen, die im Laufe des Betriebs weiter lernen – bei diesen kann man sich nicht auf eine Bewertung unmittelbar nach der Entwicklung verlassen. Das System kann durch das Lernen nicht nur besser werden, sondern auch falsche Entscheidungen lernen („model drift“). Schließlich müssen noch Veränderungen der Eingabedaten oder Rahmenbedingungen berücksichtigt werden, die zu schlechteren Entscheidungen führen können („concept drift“). Das ist zum Beispiel der Fall wenn nach einer Gesetzesänderung die auf die vorherigen Gesetze trainierte KI jetzt falsche Empfehlungen geben würde.

Bei der Berücksichtigung der Kriterien werden in den meisten Fällen Zielkonflikte nicht ausbleiben, die thematisiert und adressiert werden müssen. So kann z. B. ein Zielkonflikt zwischen Fairness und Genauigkeit auftreten. Parameter, die aus Fairnessgründen (keine ungerechtfertigte Diskriminierung auf Basis von z. B. Geschlecht, Hautfarbe, oder Migrationshintergrund) nicht berücksichtigt werden dürfen, könnten dennoch insgesamt zu einer besseren Vorhersage führen. Auch kann ein Zielkonflikt zwischen Transparenz und Sicherheit vorliegen – ein System, das komplett „durchschaubar“ ist, kann auch leichter von außen manipuliert werden.

Der Bericht von Poretschkin und anderen (2021) gibt auf 166 Seiten einen Leitfaden zur strukturierten Identifikation KI-spezifischer Risiken, eine Anleitung zur Formulierung von Prüfkriterien sowie eine Anleitung zur strukturierten Dokumentation von technischen und organisatorischen Maßnahmen für den Einsatz von KI-Anwendungen.

Es gibt noch viele weitere Richtlinien zum Einsatz von KI. Die großen Softwareunternehmen (Google, Microsoft, Apple, etc.) geben auch Empfehlungen heraus, wie KI-Anwendungen entwickelt werden sollten und welche Kriterien sie erfüllen sollten. Die „Responsible AI Practices“ von Google empfehlen u. a., menschzentriert zu entwickeln (vgl. Abschn. 6.4), unterschiedliche Metriken für Training und Überwachungen verwenden, falls möglich, die Rohdaten auf mögliche Verzerrungen zu untersuchen (vgl. Kap. 11 KI & Ethik), die Beschränkungen

des Datensatzes und des Models zu verstehen, und schließlich zu „testen, testen, testen". Insbesondere beim maschinellen Lernen sollte man nicht erwarten, dass das maschinelle Lernen herausfindet, welche Probleme gelöst werden sollten. Es ist zu überlegen, ob maschinelles Lernen wirklich einen einzigartigen Beitrag leistet (regelbasierte Lösungen nicht unterschätzen) und zu prüfen, welche Kosten durch falsche Entscheidungen (falsch positiv/falsch negativ) entstehen können (Lovejoy & Holbrook, 2017). Insbesondere zur Vermeidung von Fehlinvestitionen bietet sich die „Wizard-of-Oz"-Methode an. Der Nutzer hat bei der Evaluation der Anwendung den Eindruck mit einem KI-System zu interagieren, in Wirklichkeit werden die intelligenten Handlungen allerdings von einem Menschen getätigt, der für den Nutzer nicht sichtbar ist (daher „Wizard-of-Oz", „Pay no attention to that man behind the curtain!"). Erst wenn die so simulierte KI-Anwendung wirklich positiv bewertet wird – z. B. eine effizientere und effektivere Bearbeitung erlaubt – wird das System entwickelt.

6.6 Besondere Anforderungen der öffentlichen Verwaltung

Die öffentliche Verwaltung unterliegt weiteren, besonderen Regelungen, u. a. was die Transparenz, Nachvollziehbarkeit und Erklärbarkeit von Entscheidungen betrifft (Gode & Franke, 2019, siehe hierbei u. a. Artikel 22 der DSGVO). Entscheidungen müssen z. B. begründbar sein – und das muss mehr sein als „die KI hat es so gesagt". Des weiteren müssen bestehende Prozesse in Behörden berücksichtigt werden. Mehr Informationen dazu im Kap. 12 KI und Recht.

6.7 Fragen an KI-Anwendungen in der öffentlichen Verwaltung

Was sind Fragen, die man sich bei KI-Anwendungen in der öffentlichen Verwaltung stellen kann? Wie kann man die Gebrauchstauglichkeit sowie die weiteren Anforderungen von KI-Anwendungen überprüfen? Siehe dazu auch Unterkapitel **7.7** und 11.7.

Wie sieht die Gebrauchstauglichkeit der Anwendung aus (Batinic & Appel, 2008, EN ISO 9241-210, 2010)?

- **Effektivität:** Können Sie mit der Anwendung Ihre Ziele erreichen?

- **Effizienz:** Ist der Aufwand im Vergleich im Ergebnis gering (z. B. wenige Korrekturen notwendig)?
- **Erlernbarkeit:** Können Sie den Umgang mit der Anwendung leicht erlernen?
- **Zufriedenstellung:** Können Sie mit der Anwendung zufriedenstellend arbeiten?

Wie gut werden die KI-Anforderungen der Anwendung erfüllt (siehe dazu auch Abschn. 7.7)?

- **Fairness:** Ist die Anwendung fair (und wie ist „fair" hier definiert, vgl. Abschn. 11.7)?
- **Autonomie & Kontrolle:** Erlaubt sie Ihnen den richtigen Grad von Autonomie und Kontrolle?
- **Transparenz:** Ist das Verhalten der Anwendung transparent?
- **Verlässlichkeit:** Können Sie sich auf die Anwendung verlassen?
- **Sicherheit:** Ist die Anwendung sicher?
- **Datenschutz:** Wird der Datenschutz gewahrt?

Auch wenn man nicht alle Fragen beantworten kann, lohnt es sich, KI-Anwendungen in der öffentlichen Verwaltung kritisch und differenziert zu hinterfragen. Werden Defizite identifiziert, können sie im Rahmen des menschzentrierten Gestaltungsprozesses überarbeitet und verbessert werden.

6.8 Ihr Beitrag bei der menschzentrierten Entwicklung von KI-Anwendungen für die öffentliche Verwaltung

Die menschzentrierte Entwicklung von KI-Anwendungen kann nur dann gelingen, wenn sich die Nutzer auch bei der Entwicklung beteiligen, sprich: Feedback geben. Hierbei müssen die Nutzer, die Feedback geben, die spätere Nutzergruppe möglichst gut abbilden (repräsentativ dafür sein).

Eine relevante Eigenschaft, die berücksichtigt werden sollte, ist, wie gerne sich die Personen mit Technik auseinandersetzen (Affinität für Technikinteraktion, ATI). Diese reicht von sehr gering (die Personen interagieren mit Technik nur, wenn sie es wirklich müssen, ihnen genügt es, wenn die Technik einfach funktioniert, und grundlegendes Wissen reicht ihnen aus) bis sehr hoch (Personen möchten Technik explorieren, möchten verstehen, wie sie funktioniert, und verbringen gerne Zeit mit der Interaktion mit Technik). Ein Risiko bei der

menschzentrierten Entwicklung ist, dass sich v. a. Personen am Entwicklungsprozess durch das Testen und Geben von Feedback beteiligen, die gerne mit Technik interagieren. Personen, die keinen Spaß oder Interesse an der Interaktion mit Technik haben, überlassen das Feedback geben diesen Personen – schließlich haben diese Spaß daran und melden sich schnell dafür. Das Problem ist dann allerdings, dass die Anwendung zwar iterativ verbessert wird, allerdings für die Mitarbeiter, die gerne mit Technik interagieren. Die Benutzeroberfläche und die Funktionen werden immer mehr so gestaltet wie Personen, *die gerne mit Technik interagieren*, sie haben möchten – und immer weniger wie Personen, die nur wollen, dass die Technik einfach funktioniert. Es kann entsprechend zu einem Matthäus-Effekt kommen: „Denn wer da hat, dem wird gegeben, dass er die Fülle habe; wer aber nicht hat, dem wird auch das genommen, was er hat." (Matthäus-Evangelium, vgl. Wessel et al., 2020).

Entsprechend ist es – auch für die eigene spätere Arbeitszufriedenheit – entscheidend, dass auch Personen mit geringer Affinität für Technikinteraktion Feedback bei der Entwicklung von Anwendungen geben. Dafür sind keine technischen Kenntnisse nötig. Mitarbeiter der öffentlichen Verwaltung sind Inhaltsexperten. Sie bringen die inhaltlich-fachliche Expertise ein und bewerten die Gestaltungslösungen, indem sie damit versuchen, ihre Sachaufgaben zu bearbeiten. Wie die Anwendung dann konkret umgesetzt wird, die technisch-kreative Expertise, das ist Aufgabe der Entwickler. In der Hinsicht greift Henry Ford's Kommentar „Wenn ich die Menschen gefragt hätte, was sie wollen, hätten sie gesagt ‚schnellere Pferde'." zu kurz. Nutzer informieren die Entwickler über Anforderungen (hier z. B. schneller zum Ziel zu kommen). Wie diese technisch erreicht werden können, ist Aufgabe der Entwickler, nicht der Nutzer.

Entsprechend, beteiligen Sie sich, wenn Sie die Chance haben. Geben Sie Feedback über den gesamten Prozess, von den konkret zu erreichenden Zielen, das Verständnis und die genaue Festlegung des Nutzungskontextes, die spezifischen Nutzungsanforderungen die Konzeption bzw. Bewertung der möglichen Umsetzungen (Gestaltungslösungen) sowie die abschließende oder gar kontinuierliche Bewertung der Software. Es lohnt sich.

6.9 Übungsfragen: Gebrauchstaugliche Entwicklung von KI-Anwendungen

Zur Überprüfung Ihres Wissensstandes können Sie die folgenden Fragen beantworten.

1. Wie ist Gebrauchstauglichkeit definiert?
2. Welche Kriterien muss Gebrauchstauglichkeit nach EN ISO 9241-210:2010 und erweitert mittels Batinic & Appel (2008) erfüllen?
3. Wie können Sie die Gebrauchstauglichkeit einer Anwendung feststellen?
4. Skizzieren Sie den menschzentrierten Gestaltungsprozess nach DIN EN ISO 9241-210 (2011).
5. In welche Phasen gliedert sich ein Entwicklungsprozess üblicherweise?
6. Welche Analysen kann man u. a. durchführen, um den Nutzungskontext besser zu verstehen? Was schaut man sich in diesen Analysen jeweils an?
7. Skizzieren Sie einmal eine Ihrer typischen Arbeitstätigkeiten (ähnlich wie es Houy et al., 2020) gemacht haben. An welchen Stellen könnte die Tätigkeit mit welchen KI-Methoden unterstützt werden? Zur Erinnerung, bei Houy et al. (2020) wurden u. a. Handschriftenerkennung/OCR Robotic-Process-Automation (RPA), Natural Language Processing (NLP), und Natural Language Generation eingesetzt.
8. Welche zehn Heuristiken sollten nach Nielsen (1994) berücksichtigt werden? Was ist mit ihnen jeweils gemeint?
9. Welche weiteren Anforderungen müssen KI-Anwendungen (nach Poretschkin et al., 2021) erfüllen, um vertrauenswürdig zu sein?

6.10 Aufgaben zum eigenen Anwendungsfall

Eine KI Anwendung arbeitet selten vollständig autonom – die Nutzer werden mit der Anwendung und ihren Ergebnisse interagieren und sie teilweise auch überwachen müssen. Dafür ist ein gebrauchstaugliche Anwendung, speziell ein Mensch-KI-Interface, notwendig.

In diesem Abschnitt wird entsprechend die Gebrauchstauglichkeit der geplanten Anwendung bewertet sowie eine mögliche Evaluation des Systems konzipiert.

Hierfür bietet es sich an, das nächste Kapitel (Kap. 7: Mensch-KI-System) vor der Bewertung der Anwendung zu berücksichtigen.

- Beschreiben Sie zuerst eine Beispiel-Situation, in welcher der Nutzer mit der KI-Anwendung interagiert, anhand derer das Zielszenario Ihres Projektes deutlich wird. Z. B. das KI-System wertet die oben genannten Daten aus und macht Vorschläge – was sieht bzw. macht der Nutzer?
- Skizzieren (zeichnen! ca. eine halbe Seite) Sie die Benutzeroberfläche Ihrer KI-Anwendung (Sie können es einfach halten, man muss nur verstehen, wie der Nutzer mit dem System interagiert).
- Bewerten Sie die Gebrauchstauglichkeit mittels der Gebrauchstauglichkeitskriterien Effektivität, Effizienz, Zufriedenstellung und Erlernbarkeit. Beachten Sie auch die Anforderungen an KI-Anforderungen wie Autonomie/Kontrolle, Transparenz und Verlässlichkeit. Stellen Sie z. B. als Tabelle dar, welches Kriterium Sie wie erfassen können (z. B. anhand welcher beobachtbarer Maße oder mit welchen Fragen). Woran würde man bei der Beispielsituation erkennen, dass das Kriterium erfüllt ist?

Literatur

Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An empirical evaluation of the system usability scale. *International Journal of Human-Computer Interaction, 24*(6), 574–594.

Batinic, B., & Appel, M. (2008). *Medienpsychologie.* Springer Medizin.

DIN Deutsches Institut für Normung e. V. (2011). *DIN EN ISO 9241-210. Ergonomie der Mensch-System-Interaktion – Teil 210: Prozess zur Gestaltung gebrauchstauglicher interaktiver Systeme.* Beuth Verlag.

Gode, A., & Franke, T. (2019). KI in der ÖV – Der Computer in Erklärungsnot? In *Tagungsband der Veranstaltung am 20. März 2019 Künstliche Intelligenz – Politische Ansätze für eine moderne Gesellschaft* (S. 21–22). opencampus.sh. http://resources.opencampus.sh/190320_KI-Tagungsband.pdf. Zugegriffen: 15. Okt. 2022.

Houy, C., Gutermuth, O., Fettke, P., & Loos, P. (2020). *Potentiale Künstlicher Intelligenz zur Unterstützung von Sachbearbeitungsprozessen im Sozialwesen* (No. 8; Berichte des NEGZ). Nationales E-Government Kompetenzzentrum e. V.

Lovejoy, J., & Holbrook, J. (2017). Human-centered machine learning. *Google Design.* https://medium.com/google-design/human-centered-machine-learning-a770d10562cd. Zugegriffen: 15. Okt. 2022.

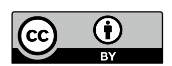
Nielsen, J. (1994/2020). 10 Usability heuristics for user interface design. *NN/g Nielsen Norman Group.* https://www.nngroup.com/articles/ten-usability-heuristics/. Zugegriffen: 15. Okt. 2022.

Poretschkin, M., Schmitz, A., Akila, M., Adilova, L., Becker, D., Cremers, A., Hecker, D., Houben, S., Mock, M., Rosenzweig, J., Sicking, J., Schulz, E., Voss, A., & Wrobel, S. (2021). *Leitfaden zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz – KI-Prüfkatalog*. Fraunhofer IAIS. www.iais.fraunhofer.de/ki-pruefkatalog. Zugegriffen: 15. Okt. 2022.

Responsible AI practices. (n.d.). *Google AI*. https://ai.google/responsibilities/responsible-ai-practices/. Zugegriffen: 15. Okt. 2022.

Wessel, D., Heine, M., Attig, C., & Franke, T. (2020, September). Affinity for technology interaction and fields of study – Implications for human-centered design of applications for public administration. *Mensch und Computer 2020 (MuC'20)*. https://doi.org/10.1145/3404983.3410020.

Zhu, Y.-Q., Corbett, J., & Chiu, Y.-T. (2021). Understanding employees' responses to artificial intelligence. *Organizational Dynamics, 50*(2), 100786. https://doi.org/10.1016/j.orgdyn.2020.100786.



7 Mensch-KI-System

Zusammenfassung

Bei KI-Anwendungen arbeiten Mensch und KI-System eng zusammen. Diese Zusammenarbeit muss entsprechend gestaltet sein. In diesem Kapitel werden Formen der Zusammenarbeit vorgestellt und die Rolle von KI im Arbeitsalltag hinterfragt. Grade der Automation und Kriterien guter Zusammenarbeit werden vorgestellt und Hinweise für die konkrete Gestaltung der Zusammenarbeit gegeben. Zur Bewertung der Qualität der Zusammenarbeit werden Fragen an KI-Anwendungen vorgestellt.

7.1 Einleitung

Wenn Mensch und KI-System zusammenarbeiten, stellt sich die Frage, wie diese Zusammenarbeit bestmöglich gestaltet werden kann. Was macht eine gute Zusammenarbeit aus? Worauf muss geachtet werden? Dafür werden zuerst drei Fallbeispiele von guter, schlechter und hässlicher Zusammenarbeit (Abschn. 7.2), unterschiedliche Arten der Zusammenarbeit (7.3), speziell wie KI-Systeme im Arbeitskontext gesehen werden, und verschiedene Grade der Automation (7.4) dargestellt. Dann werden Kriterien guter Zusammenarbeit in Mensch-KI-Systemen (7.5) sowie zu der konkreten Gestaltung der Zusammenarbeit (7.6) vorgestellt, was das vorherige Kap. 6 (Gebrauchstaugliche Entwicklung von KI-Anwendungen) erweitert. Abschließend folgen Fragen zur Bewertung von KI-Anwendungen (7.7), Übungsfragen (7.8) sowie Aufgaben zum eigenen Anwendungsfall (7.9).

M. Heine et al., *Künstliche Intelligenz in öffentlichen Verwaltungen*, Edition eGov-Campus, https://doi.org/10.1007/978-3-658-40101-6_7

7.2 Fallbeispiele

Betrachtet man KI-Anwendungen daraufhin, wie gut Mensch und KI zusammenarbeiten, lassen sich gute, schlechte und hässliche Anwendungen identifizieren.

Gute Anwendungen erhöhen die Leistung des Menschen. Man spricht hier von „augmented intelligence". Mensch und KI erreichen zusammen Leistungen, die ein Mensch oder die KI alleine nicht erreicht hätte. Ein anschauliches Beispiel ist Advanced Chess (auch Cyborg Chess oder Centaur Chess). Wenn Mensch und KI-System als ein Spielpartner zusammenarbeiten, sind sie einem Menschen oder einem KI-System alleine überlegen.

Schlechte Anwendungen vermindern die Leistung des Menschen. Ein bekanntes Beispiel ist der in Kap. 6 schon erwähnte „Clippy" (Karl Klammer) von Microsoft Word. Dieses Hilfesystem sollte den Nutzer beim Schreiben unterstützen, unterbrach ihn aber bei der Arbeit – insbesondere durch die Animation, welche Aufmerksamkeit auf sich zieht. In diesem Beispiel konnte man diese „Tipps" ausstellen, um zumindest mit normaler Leistung weiter arbeiten zu können.

Hässliche Anwendungen verursachen durch die schlechte Zusammenarbeit zwischen Mensch und KI-System nicht nur eine geringere Leistung als ohne KI, sie führen auch zu gravierenden Schäden, die ohne KI nicht aufgetreten wären. Ein sehr negatives Beispiel ist hier das Maneuvering Characteristics Augmentation System (MCAS) der Boeing 737 MAX. Das MCAS sollte eigentlich dabei helfen, das Flugzeug zu stabilisieren. Da die Piloten unter anderem nicht ausreichend über das System informiert waren, kam es zu zwei Flugzeugabstürzen mit über 300 Toten. Dies ist auch ein Beispiel für „algorithmic hubris", der Versuch von Programmierern, „narrensichere" Systeme („foolproof systems") zu entwickeln (Shneiderman, 2020a). Ein Beispiel aus dem Verwaltungskontext ist ein KI-System, das fälschlicherweise massenhaft überhöhte Steuerbescheide verschickt, die bei den Bürgern nicht nur Verunsicherungen, Frustration und Aggression auslösen, sondern auch Existenzen infrage stellen können (Rohde, 2017).

Die Frage ist jetzt, wie die Zusammenarbeit im Mensch-KI-System gestaltet sein muss, damit Mensch und KI sich wie beim Advanced Chess ergänzen und bessere Leistung als alleine erbringen können und Störungen wie bei Clippy oder Katastrophen wie beim MCAS verhindert werden.

7.3 Arten der Zusammenarbeit

Wie kann eine Zusammenarbeit zwischen Mensch und KI gestaltet sein, speziell, wie wird die KI im Arbeitsalltag gesehen?

Ein relatives altes, aber immer noch nützliches Modell ist das MABA-MABA („Men Are Better At – Machines Are Better At“, oder auch HABA-MABA, „Humans Are Better At – Machines Are Better At“) Modell von Fitts (1951). Menschen und Maschinen haben ihre jeweiligen Stärken und je nachdem, wer in der konkreten Tätigkeit besser ist, übernimmt diese Tätigkeit. So sind Menschen z. B. besser im Fällen von Urteilen, Induktion, und Improvisation, während Maschinen schneller sind, hochkomplexe Operationen ausführen können und sehr gut parallel arbeiten können.

Zwar haben seit 1951 Computer in vielen Bereichen aufgeholt und auch heute verschieben sich noch Bereiche, in denen Maschinen Menschen übertreffen. Auch ist die Arbeitstätigkeit üblicherweise eng verzahnt, was dazu führt, dass Mensch und Maschine in vielen Arbeitsschritten eng zusammenarbeiten müssen.

Eine andere Sichtweise, die auch in Folge der zunehmenden Leistung und höherer Automation (siehe Abschn. 7.4) vermutlich häufiger auftreten wird, ist „die KI“ vermenschlicht (anthropomorphisiert) als Kollegin zu betrachten. Man arbeitet mit „der Kollegin“ KI zusammen und delegiert die Tätigkeit an diese. „Sie“ macht die Aufgaben. Ein Problem dabei ist, dass die Kontrolle der und die Verantwortung für die Arbeitstätigkeit weiterhin beim Menschen liegen muss – nie bei der KI (siehe dazu auch Kap. 11, KI & Ethik). Das KI-System kann keine Verantwortung übernehmen – überspitzt gesagt kann man einem frustrierten Bürger bei Fehlern nicht sagen: „Sie war’s!“. Entsprechend sollte das Konzept der „Kollegin“ kritisch hinterfragt werden (vgl. Shneiderman, 2020a).

Hilfreicher ist eine Metapher von Steve Jobs, damals noch über Computer selbst, nicht speziell zu KI-Systemen. In einem Interview verwies er auf einen Artikel, der die menschliche Leistung bei der Fortbewegung mit der von Tieren verglich. Menschen wurden dabei von diversen Tieren, was den Energieaufwand pro Streckeneinheit betrifft, weit geschlagen (z. B. braucht der Kondor am wenigsten Energie). Die Autoren des Artikels sahen sich dann an, was passiert, wenn der Mensch ein Fahrrad verwendet. Konsequenz – der Mensch schlug den Kondor bei weitem. Jobs übertrug diesen Vergleich auf Computer als er sagte: „What a computer is to me is the most remarkable tool that we have ever come up with. It’s the equivalent of a bicycle for our minds.“ (Übersetzt: „Was ein Computer für mich ist, das ist das bemerkenswerteste Werkzeug, das wir jemals entwickelt haben. Es ist das Äquivalent eines Fahrrads für den menschlichen Verstand.“). Auch wenn er über Computer generell gesprochen hat, das Ziel

von KI sollte sein, die menschliche Leistung zu unterstützen: zum Beispiel die Arbeitsziele mit weniger Aufwand (höhere Effizienz) zu erreichen oder zu einer Entlastung zu führen, damit man sich auf die wichtigen Dinge konzentrieren kann. KI soll den Arbeitsalltag erleichtern, dem Menschen erlauben, Tätigkeiten auszuführen, zu denen er sonst so nicht fähig wäre. Vergleichbar mit dem Fallbeispiel des Advanced Chess (7.2) haben KI-Systeme das Potenzial(!) menschliche Grenzen zu überwinden und zu Ergebnissen zu führen, die von einem Menschen alleine nicht, oder nicht so effizient, erreicht werden können.

7.4 Automation

Bei KI-Systemen wird häufig der Begriff Automation verwendet – Automation meint dabei die Übernahme von Funktionen eines Prozesses durch künstliche Systeme, wobei insbesondere auch Steuerungsaufgaben einbezogen werden (Voigt, 2018). Wie kann man sich diese Automation vorstellen – welche unterschiedlichen Formen gibt es dabei?

Ein älteres aber immer noch nützliches Modell stammt von Sheridan und Verplank (1978). Es stellt die zunehmende Automation auf einer Dimension in zehn Stufen dar, von „der Mensch führt die gesamte Tätigkeit aus bis er sie an den Computer übergibt“, bis hin zu „der Computer führt die gesamte Tätigkeit aus, sofern er entscheidet, dass sie durchgeführt werden sollte, und entscheidet, ob Nutzer informiert wird“. In den Zwischenstufen hilft der Computer zu unterschiedlichen Graden, schlägt Aktionen vor, oder führt sie aus.

Bei der Betrachtung konkreter Arbeitstätigkeiten stellen sich folgende Fragen: Welcher Grad der Automation wäre für die entsprechende Tätigkeit oder Teilaufgabe dieser Tätigkeit akzeptabel – und warum? Wie könnte diese Automationsstufe für die jeweilige Tätigkeit erreicht werden?

Die Automationsgrade lassen sich auch einfacher zusammenfassen (Poretschkin et al., 2021) – von Human Control (maximal Vorschläge des Computers), Human-in-the-Loop (Mensch muss Vorschläge/Entscheidungen der KI vor Ausführung bewilligen), Human-on-the-Loop (KI arbeitet normalerweise autonom, aber Mensch kann korrigierend eingreifen), bis Human-out-of-the-Loop (KI arbeitet autonom, kann sie höchstens deaktivieren).

Während Sheridan und Verplank (1978) und auch die Autonomiegrade von Human Control bis Human-out-of-the-Loop Automation als eindimensional sehen (von „der Mensch macht die gesamte Tätigkeit“ zu „der Computer führt die

gesamte Tätigkeit aus und entscheidet sogar selbst darüber, ob der Nutzer informiert wird“), gibt es auch Modelle, die menschliche Kontrolle und Automation des Computers als zwei getrennte Dimensionen sehen.

Das Human-Centered Artificial Intelligence (HCAI) Framework von Shneiderman (2020a) ist ein solches Rahmenmodell. Es soll ermöglichen, verlässliche, sichere und vertrauenswürdige KI-Anwendungen zu entwickeln und sieht menschliche Kontrolle vs. Automation nicht als eindimensional, sondern als zwei getrennte Dimensionen. Damit soll erreicht werden, sowohl einen hohen Grad an menschlicher Kontrolle und ein hoher Grad an Automation zu erlauben (sofern notwendig), als auch zu verstehen, wann die vollständige Kontrolle von Mensch oder Computer notwendig ist, und die Gefahren exzessiver Kontrolle von Mensch oder Computer zu vermeiden.

Je nach Tätigkeit sind unterschiedliche Grade von menschlicher Kontrolle und Automation durch den Computer notwendig. Das Ziel ist hierbei nicht „je mehr Kontrolle oder je mehr Autonomie desto besser“, sondern das richtige Maß auf beiden Dimensionen zu finden. Beispiele aus der öffentlichen Verwaltung wären:

- bei hoher Computerkontrolle und geringer menschlicher Kontrolle die Optische Zeichenerkennung (OCR),
- bei hoher menschlicher Kontrolle und geringer Computerkontrolle Policy-Entscheidungen, die mit anderen Stakeholdern ausdiskutiert werden müssen und erst in diesem menschlichen Zusammenspiel entwickelt werden, sowie bei
- „verlässliche, sichere und vertrauenswürdige KI“ (beide Dimensionen hoch ausgeprägt), Assistenzsysteme, welche die Nutzer im richtigen Zeitpunkt im richtigen Ausmaß unterstützen.

Das HCAI-Modell ist interessant, um Automation nicht als eine Dimension zu sehen, sondern bewusst nach Möglichkeiten zu suchen, bei denen – wenn hilfreich – sowohl menschliche Kontrolle als auch Computer Automation das richtige Maß aufweisen.

Abschließend sollte bei Automation das Problem des richtigen Grades an Vertrauen in die Automation nicht unterschätzt werden. Man kann der KI zu sehr vertrauen („overtrust“) und dabei Fehler des KI-Systems übersehen (z. B. indem Entscheidungen unkritisch akzeptiert werden, insbesondere wenn das KI-System „eigentlich immer“ richtige Entscheidungen getroffen hat). Man kann dem System aber auch zu wenig vertrauen („undertrust“) und es häufig aber auch unnötigerweise überwachen. Insbesondere die Überwachung von automatisierten

Tätigkeiten erfordert konstante Aufmerksamkeit (Vigilanz) und kann langfristig anstrengender sein, als die Tätigkeit selbst zu durchzuführen. Entsprechend muss der Nutzer richtig einschätzen können, unter welchen Bedingungen das KI-System welche Leistung zeigt.

7.5 Kriterien guter Zusammenarbeit in Mensch-KI-Systemen

Welche Kriterien muss ein Mensch-KI-System erfüllen, damit Mensch und KI gut zusammenarbeiten können? Eine notwendige Vorbedingung ist, dass der Nutzer über den Einsatz von KI immer informiert ist („Sichtbare KI", Abschn. 7.5.1). Dann machen Autonomie und Kontrolle (Abschn. 7.5.2), Transparenz/Nachvollziehbarkeit (Abschn. 7.5.3), Verlässlichkeit (Abschn. 7.5.4), und Sicherheit (Abschn. 7.5.6) einen großen Teil der Vertrauenswürdigkeit eines KI-Systems aus (Poretschkin et al., 2021) und sind damit für eine gute Zusammenarbeit zentral. Es gibt auch weitere Rahmenmodelle, die ähnliche Kriterien postulieren (Abschn. 7.5.7).

7.5.1 Vorbedingung: Verwendung von KI offen legen

Der Nutzer muss immer wissen, wenn er mit einem KI-System interagiert. Das war u. a. auch ein Problem bei der Boeing 737 MAX mit dessen Maneuvering Characteristics Augmentation System (MCAS, vgl. Abschn. 7.2). Die Piloten waren u. a. nicht ausreichend informiert, was das System macht. IBM's (2019) „Everyday Ethics for Artificial Intelligence" bringt es auf den Punkt mit: „Your users should always be aware that they are interacting with an AI. Good design does not sacrifice transparency in creating a seamless experience. Imperceptible AI is not ethical AI." [„Ihren Nutzern sollte immer bewusst sein, dass sie mit einer KI interagieren. Gutes Design opfert nicht die Transparenz, um eine nahtlose Erfahrung zu erzeugen. Nichtwahrnehmbare KI ist keine ethische KI."].

7.5.2 Autonomie und Kontrolle

Bei Autonomie und Kontrolle muss der richtige Grad an Autonomie für die Anwendung gewählt werden (Human-in/on/out-of-the-Loop, siehe Abschn. 7.4) und der Mensch durch die KI-Anwendung angemessen unterstützt werden

(Poretschkin et al., 2021). Insbesondere muss ausreichend Handlungsspielraum des Menschen bei der Verwendung des KI-Systems zur Verfügung stehen.

Hierbei muss (u. a. nach Poretschkin et al., 2021) der Vorrang des menschlichen Handelns gewährleistet werden (informierte, bewusste Abgabe an das KI-System). Apple's Guidelines bringen es mit „Menschen, nicht Apps, haben die Kontrolle“ (Apple, 2022) auf den Punkt. Des Weiteren muss eine angemessene und verantwortungsvolle Gestaltung der Aufgabenverteilung zwischen Mensch und KI-Anwendung hergestellt werden und die Nutzer (und Betroffen!) müssen über die Verwendung informiert sein, die Nutzer müssen das KI-System bedienen können und die Kontrolle notfalls auch zurückholen können (ggfs. über das Abschalten der KI).

Nach Poretschkin et al. (2021) kann man Autonomie und Kontrolle u. a. dadurch gewährleisten, dass man die relevanten Personengruppen und Organisationen in die Entwicklung einbindet, konsequent auf den Vorrang menschlichen Handelns achtet, wirksame Beschwerdemöglichkeiten etabliert, ein klares Rollen- und Rechtekonzept für die Nutzung der KI-Anwendung etabliert, die KI-Anwendung unter menschlicher Aufsicht stellt und Abschalt-Szenarien sowohl definiert als auch technisch möglich macht. Im Falle eines Abschaltens einer KI-Anwendung bedeutet dies allerdings auch, dass der frühere Nutzer weiterhin in der Lage sein muss, die Tätigkeit auszuführen (wenn auch nicht so effizient). Die Informiertheit und Befähigung von Nutzern und Betroffenen muss entsprechend weiterhin gewährleistet werden. Dazu gehören nach Poretschkin et al. (2021) u. a. die ausreichende Qualifikation der Nutzer sowie die Sicherstellung der Vollständigkeit, Sichtbarkeit und Zugänglichkeit der Informationen im Abschalt-Szenario. Diese Informationen dürfen z. B. nicht nur im laufenden KI-System verfügbar sein, sonst wird mit dem Abschalten auch die Datenbasis für eine Entscheidung ohne KI genommen.

7.5.3 Transparenz/Nachvollziehbarkeit

Ein weiteres zentrales Kriterium bei Automation ist Transparenz bzw. Nachvollziehbarkeit der Entscheidungen. Man würde z. B. in der Zusammenarbeit mit einem Kollegen bei der Frage, warum er eine bestimmte Entscheidung getroffen hat, ein „glaub' es mir einfach, dass das stimmt“ kaum akzeptieren. Bei einem KI-System sollte es nicht anders sein. IBM's (2019) Everyday Ethics for Artificial Intelligence gibt die Empfehlung: „AI should be designed for humans to easily perceive, detect, and understand its decision process. In general, we don't blindly trust those who can't explain their reasoning. The same goes for AI, perhaps

even more so. As an AI increases in capabilities and achieves a greater range of impact, its decision-making process should be explainable in terms people can understand.“ [KI sollte so entworfen werden, dass es für Menschen einfach ist, den Entscheidungsprozess wahrzunehmen, zu erkennen und zu verstehen. Im Allgemeinen vertrauen Menschen anderen Personen nicht blind, wenn diese ihre Schlussfolgerungen nicht erklären können. Das gleiche gilt für KI, vielleicht noch mehr. Wenn eine KI mehr Fähigkeiten bekommt und einen größeren Einflussbereich erreicht, sollte der Entscheidungsprozess in Begriffen erklärbar sein, die Menschen verstehen können.]

Zur Transparenz und Nachvollziehbarkeit gehören nach Poretschkin et al. (2021) u. a. die Erklärbarkeit, wie die Vorhersage zustande gekommen ist, die Interpretierbarkeit des Modells, z. B. dass das verwendete maschinelle Lernverfahren als Ganzes transparent ist, und die Nachverfolgbarkeit und Reproduzierbarkeit von Entscheidungen. Letzteres ist u. a. für rechtliche Fragen relevant und kann z. B. mittels Logdaten, Dokumentationen bzw. Archivierungen des Designs, der Daten, des Trainings, des Testens und Validierens des Modells erreicht werden.

Bei der Transparenz und Nachvollziehbarkeit muss unterschieden werden zwischen Nutzern, bei denen es v. a. um die sichere, ordnungsgemäße, verantwortungsvolle Bedienung geht, und KI-Experten, die sich z. B. mit dem Aufdecken von Modellschwächen beschäftigen. Mitarbeiter öffentlicher Verwaltungen müssen keine KI-Experten werden. Sie sollten aber über Handlungswissen verfügen, um z. B. die Qualität von Entscheidungen einzuschätzen und mögliche Probleme zu erkennen.

Transparenz und Erklärbarkeit sind je nach Umsetzung der KI-Anwendung unterschiedlich gut möglich (Gode & Franke, 2019). Während bei Whitebox- oder Glassbox-Systemen, wie z. B. regelbasierten Entscheidungssystemen, die Regeln direkt überprüfbar sind, kann man bei Blackbox-Systemen, wie z. B. der Texterkennung bei maschineller Dokumenterfassung, eine Erklärbarkeit nur sehr schwierig herstellen. Entsprechend sollten diese Systeme nur bei Tätigkeiten eingesetzt werden, bei denen eine Erklärbarkeit im Detail nicht notwendig ist und die Qualität anhand der Ergebnisse überprüfbar ist (wie z. B. bei der Texterkennung). Mehr Informationen zum Thema Erklärbarkeit gibt es in Kap. 8 („Erklärbare KI“).

7.5.4 Verlässlichkeit

Die Verlässlichkeit eines KI-Systems umfasst nach Poretschkin et al. (2021) u. a. die Korrektheit der Ausgaben, Angaben zur Einschätzung der Modellunsicherheit beim maschinellen Lernen, die Robustheit gegenüber gestörten oder manipulierten Eingaben, den Umgang mit unerwarteten Situationen, das Wissen über die Grenzen des Modells sowie das Abfangen von Fehlern. Verlässlichkeit ist bei jedem System mindestens teilweise relevant. Wäre die Verlässlichkeit nicht relevant, dann könnte ein System auch einfach Zufallsentscheidungen treffen (Poretschkin et al., 2021).

Ein zentraler Punkt bei Verlässlichkeit ist die Kommunikation von Unsicherheit. Wie sehr kann sich der Nutzer auf das System bzw. eine bestimmte Entscheidung verlassen? Entscheidungen, die das System nicht mit ausreichender Sicherheit tätigen kann, müssen klar kommuniziert werden. Allerdings ist die Feststellung, wie sicher das Ergebnis eines Systems ist, nicht trivial. Man kann sich zwar leicht vorstellen, dass das KI-System einen Prozentwert bezüglich der Sicherheit zurückgibt (oder als Icon einen „Daumen hoch“), aber die Frage dabei ist, wie kommt das System auf diesen Wert (oder den „Daumen hoch“)? Anhand welcher Kriterien erfolgt diese Bewertung? Um eine solche Einschätzung umzusetzen, benötigt man u. a. umfangreiches Domänenwissen von Mitarbeitern der öffentlichen Verwaltung, welche die Vorgänge sehr gut kennen. Hinzu kommt dann mathematisch-technische Expertise von KI-Experten, welche die KI-Anwendung selbst entwickeln bzw. trainieren.

Ein Beispiel ist ein KI-basiertes Übersetzungsprogramm (Poretschkin et al., 2021). Die Frage ist hier: Wie gut ist die Übersetzung? Ideal wäre ein Wert, der die Qualität der Übersetzung angibt, zum Beispiel der BLEU-Wert (bilingual evaluation understudy score). Aber auch hier gibt es Abmessungsentscheidungen. Ist es die richtige Metrik und was ist der richtige Schwellenwert, der überschritten werden muss?

7.5.5 Robustheit

Wenn bei der Verlässlichkeit von der Robustheit gesprochen wird, dann geht es nach Poretschkin et al. (2021) u. a. um den Umgang mit kleineren Störungen (z. B. Bildverzerrungen, Sensorrauschen/-ausfall oder unpräzise Datenerhebung wie Mess- oder Tippfehler) und adversarialen Fällen (kleine Abweichung mit großer Wirkung, falls absichtlich eingesetzt auch „adversariale Attacke“).

Die Robustheit eines Systems kann sich im Laufe des Betriebs verändern (Poretschkin et al., 2021). Das kann über Model Drift passieren, falls das System weiter lernt und im Laufe des Lernprozesses an Verlässlichkeit einbüßt, oder Concept Drift, wenn sich der Anwendungskontext oder die äußeren Bedingungen ändern (z. B. über Gesetzesänderungen). Insbesondere der Concept Drift sollte bei der Entwicklung eingeplant sein, sonst hat man zwar ein KI-System, kann es aber unmodifiziert nicht mehr weiter einsetzen.

Um die Robustheit zu gewährleisten muss nach Poretschkin et al. (2021) u. a. der Anwendungsbereich klar definiert sein (beim maschinellen Lernen müssen die Trainingsdaten diesen abdecken), eine klare Operationalisierung der Anforderungen erfolgen (wie wird es gemessen?) und das Modell mit „herausfordernden Eingabedaten“ (sogenannte „Corner Cases“) getestet werden. Außerdem sollten „Sanity Checks“ eingeplant sein (in welchen Bereichen müssen Daten bleiben, z. B. bei der Texterkennung eines handschriftlich ausgefüllten Formulars wäre das Alter einer Person kleiner als 0 Jahre oder älter als 120 Jahre sehr unwahrscheinlich).

7.5.6 Sicherheit

Bei der Sicherheit unterscheidet man nach Poretschkin et al. (2021) zwischen der funktionalen Sicherheit („Safety“) und der IT-Sicherheit („Security“).

Bezüglich der funktionalen Sicherheit („Safety“) geht es v. a. um den Schutz der Außenwelt vor einem funktionalen Versagen des KI-Systems. Ein klassisches KI-Beispiel ist der Schutz der Fußgänger vor Unfällen beim autonomen Fahren. In der öffentlichen Verwaltung ist das Verhindern vom massenhaft automatisiert ausgesendeten falschen Mahnungen ein eindrückliches Beispiel (vgl. Abschn. 7.2). Hierbei können u. a. Sanity Checks (ist es realistisch, wenn plötzlich sehr viele Personen hohe Nachzahlungsaufforderungen erhalten?) und Fail-Safe States wie ein Abschalten der KI helfen.

Bezüglich der IT-Sicherheit („Security“) geht es v. a. um die Integrität und Verfügbarkeit der Anwendung. Integrität meint den Schutz des KI-Systems vor der Umgebung (z. B. Angriffe, inkl. via gezielte Manipulation der Datenbasis, sprich „Data Poisoning“). Die Verfügbarkeit kann nicht nur aufgrund von einem technischen Hardware-Ausfall infrage gestellt werden. Durch externe Angriffe kann ein System auch ganz oder teilweise nicht mehr nutzbar sein (z. B. durch Denial-of-Service-Attacken, bei denen extrem viele Anfragen das KI-System überlasten). Gerade im Bereich von KI-Systemen, die weiter lernen, kann das

System aber auch seine Funktion verlieren und dadurch nicht mehr verfügbar, d. h. einsetzbar, sein.

Insbesondere die Verfügbarkeit sollte man nicht unterschätzen. Gerade beim maschinellen Weiterlernen kann ein System kompromittiert werden. Ein Beispiel sind Chatbots, die in der Interaktion mit den Benutzern dazulernen. So wurde Microsoft's Chatbot „Tay" auf Twitter innerhalb von weniger als 24 Stunden zu Aussagen bewegt, die dazu geführt haben, dass Microsoft den Chatbot vom Netz genommen hat. Tay sollte von Unterhaltungen lernen und dadurch immer besser werden. Sie wurde allerdings vor allem von Online-Trollen mit Aussagen „gefüttert", die dazu geführt haben, dass sie rassistische Verunglimpfungen bis hin zu Aufrufen zu Genozid von sich gegeben hat. Letztendlich wurde der Chatbot von den Betreibern vom Netz genommen, was ein PR-Desaster für Microsoft war. In der öffentlichen Verwaltung würde man solche Chatbots derzeit noch nicht einsetzen, da die dort eingesetzten Chatbots auf Basis einer festen und nicht vom Nutzer veränderlichen Wissensbasis operieren. Es zeigt allerdings die Gefahren eines solchen „selbstlernenden Systems", wenn nicht kontrolliert werden kann, von wem es dazulernt. Es ist dann nicht mehr verfügbar und die bisher vom System übernommene Tätigkeit (z. B. Auskünfte geben) würden wieder von den Mitarbeitern übernommen werden müssen.

7.5.7 Weitere Rahmenmodelle

Neben diesen Kriterien aus dem KI-Prüfkatalog von Poretschkin et al. (2021) gibt es weitere Modelle, wie ein Mensch-KI-System gestaltet sein sollte. Das Human-Centered Artificial Intelligence (HCAI) Framework von Shneiderman (2020b) sieht dabei z. B. Verlässlichkeit (Audits, Dokumentation, Analyse-Werkzeuge, Benchmark Tests, kontinuierliche Begutachtung der Datenqualität und Testen auf mögliche Verzerrungen, Design-Strategien die Vertrauen schaffen, Erklärbare KI-Ansichten), Sicherheit (Verpflichtung zur Sicherheit durch Führungskräfte, offenes Berichten über Fehler und kritische Ereignisse, öffentliche Berichte von Problemen und zukünftigen Plänen) und Vertrauenswürdigkeit (Einhalten von Standards und Richtlinien, Zertifizierung, externe Kontrolle) als wichtige Kriterien.

Das Rahmenmodell geht dabei über das hinaus, was ein individueller Nutzer leisten kann, und setzt auch einen entsprechenden Umgang im Team, in der Organisation und in der Industrie selbst (hier: öffentliche Verwaltung) voraus (Shneiderman, 2021). Dennoch lohnt es sich, die Kriterien guter Zusammenarbeit in Mensch-KI-Systemen auch als Nutzer zu betrachten.

7.6 Gestaltung der Zusammenarbeit in Mensch-KI-Systemen

Worauf ist bei der Entwicklung von KI-Anwendung zu achten, damit die Zusammenarbeit zwischen Mensch und System gut funktioniert? Dieses Unterkapitel erweitert das Kap. 6 („Gebrauchstauglichen Entwicklung von KI Anwendungen“), hier allerdings stärker auf den Interaktionsaspekt mit dem KI-System.

Die großen Softwareunternehmen (Microsoft, Apple, Google) haben Richtlinien für Mensch-KI-Interaktion herausgegeben. In diesem Unterkapitel stehen die Richtlinien von Microsoft (Amershi et al., 2019; Microsoft, 2019) im Vordergrund, da sie eine hilfreiche Übersicht darstellen und klar nach Phasen gegliedert sind:

- Zu Beginn deutlich machen, was das System kann und wie gut das System dies machen kann.
- Während der Interaktion den Kontext (Aufgabe, Umgebung) berücksichtigen, kontextrelevante Informationen zeigen, relevante soziale Normen berücksichtigen und soziale Voreingenommenheiten abmindern (siehe dazu auch Kap. 11: KI & Ethik).
- Bei Fehlern, und hier wird realistischerweise davon ausgegangen, dass Fehler passieren und der Nutzer gut damit umgehen sollte, den effizienten Aufruf und das effiziente Beenden des KI-Systems unterstützen, effiziente Korrektur unterstützen, im Zweifel den Handlungsspielraum des Dienstes verändern (z. B. das System registriert eine hohe Unsicherheit bei einer Entscheidung und gibt eine Bitte um eine Nutzerentscheidung anstatt die Entscheidung selbst durchzuführen), und deutlich machen, warum das System das gemacht hat, was es gemacht hat.
- Über die Zeit sollte das System sich an den Nutzer anpassen, indem es die letzten Interaktionen erinnert, vom Nutzungsverhalten lernt, Updates und Anpassungen behutsam vornimmt, den Nutzer zu Feedback anregt, die Konsequenzen des Nutzerverhaltens verdeutlicht, eine globale Kontrolle erlaubt, sowie den Nutzer über Veränderungen informiert.

Es gibt allerdings diverse weitere Gestaltungsrichtlinien. Auch Shneiderman (2020a, „Human-Centered Artificial Intelligence (HCAI) Framework“, vgl. Abschn. 7.4) hat z. B. „Prometheus Prinzipien“ aufgestellt. Dazu gehören eine konsistente Benutzeroberfläche, die es Nutzern erlaubt, Absichten zu formen, auszudrücken und zu widerrufen, das kontinuierliche Zeigen der interessanten

Objekte und Aktionen, schnelle, inkrementelle und reversible Aktionen, die Prävention von Fehlern, informatives Feedback um jede Aktion des Nutzers zu bestätigen, Fortschrittsanzeigen, und Berichte über abgeschlossene Handlungen.

Es lohnt sich, zu prüfen inwiefern derzeit bekannte KI-Systeme diesen Kriterien genügen (die von Amershi et al., 2019 und die von Shneiderman, 2020a).

7.7 Fragen an KI-Anwendungen in der öffentlichen Verwaltung

Was sind Fragen, die man sich bei KI-Anwendungen in der öffentlichen Verwaltung stellen kann? Wie kann man die Gebrauchstauglichkeit sowie die weiteren Anforderungen von KI-Anwendungen überprüfen? Siehe dazu auch Unterkapitel 6.7 und 11.7.

Mensch-KI-Interaktion (Amershi et al., 2019**, kombiniert mit** Shneiderman, 2020a**).**

Zu Beginn

- Macht die KI-Anwendung deutlich, was sie kann?
- Macht die KI-Anwendung deutlich, wie gut sie es machen kann?

Während der Interaktion

- Erlaubt die KI-Anwendung (v. a. deren Benutzeroberfläche) es Ihnen zu überlegen, was Sie erreichen möchten, die Absichten auch umzusetzen und ggfs. auch rückgängig zu machen?
- Berücksichtigt die KI-Anwendung den Kontext (Aufgabe/Umgebung)?
- Zeigt die KI-Anwendung kontextrelevante Informationen?
- Sind die für Sie relevanten Informationen und Handlungen der KI-Anwendung kontinuierlich für Sie sichtbar?
- Erhalten Sie informatives Feedback, wenn Sie die Anwendung bedienen (Eingaben und andere Aktionen durchführen)?
- Wird Ihnen der Fortschritt der KI-Anwendung angezeigt?
- Erhalten Sie einen Bericht über abgeschlossene Handlungen?
- Berücksichtigt die KI-Anwendung relevante soziale Normen?
- Mindert die KI-Anwendung soziale Voreingenommenheiten?

- Werden Fehler durch die Anwendung soweit wie möglich verhindert (z. B. indem keine ungültigen Eingaben möglich sind)?

Bei Fehlern

- Können Sie die KI-Anwendung mit wenig Aufwand aufrufen (schnelle, inkrementelle und reversible Aktionen)?
- Können Sie die KI-Anwendung mit wenig Aufwand beenden?
- Können Sie Korrekturen mit wenig Aufwand durchführen?
- Reduziert die KI-Anwendung bei Unsicherheit ihren Handlungsspielraum (z. B. Hinweis auf Auffälligkeit statt Autokorrektur)?
- Macht die KI-Anwendung deutlich, warum sie gemacht hat, was sie gemacht hat?

Über die Zeit

- Erinnert sich die KI-Anwendung an die letzten Interaktionen?
- Lernt die KI-Anwendung von Ihrem Verhalten?
- Werden Updates und Anpassungen behutsam durchgeführt (Updates führen nicht zu gravierenden Veränderungen)?
- Werden Sie zum Feedback angeregt?
- Werden Ihnen die Konsequenzen Ihres Nutzerverhaltens verdeutlicht?
- Können Sie die KI-Anwendung global kontrollieren (Einstellungen an einer Stelle, die sich auf das gesamte Verhalten des Systems auswirken)?
- Werden Sie von der KI-Anwendung über Veränderungen informiert (bei Updates z. B. bezüglich neuer/veränderter Fähigkeiten der KI-Anwendung)?

KI-Anforderungen (Poretschkin et al., 2021)

Fairness

- Ist die Anwendung fair – nach welcher Definition von Fairness?
- Werden unverzerrte, faire, Entscheidungen getroffen? (siehe dazu Abschn. 11.7)

Autonomie/Kontrolle

- Erlaubt das KI-System Ihnen einen angemessenen Grad von Autonomie und Kontrolle?

- Hat menschliches Handeln weiterhin Vorrang (informiert, bewusste Abgabe an KI)?
- Ist die Automationsstufe passend (Automation z. B. via Stufen von Sheridan & Verplank, 1978; Human Control/Human-in/on/out-of-the-Loop; Shneiderman, 2020a)?
- Haben Sie ausreichend Handlungsspielraum?
- Gibt es einen Ermessensspielraum, der berücksichtigt sein muss – wird dieser auch berücksichtigt?
- Können Sie das Vorgehen der KI-Anwendung kontrollieren?
- Können Sie die Ergebnisse der KI-Anwendung überprüfen?
- Sind Sie weiterhin informiert (vollständige, sichtbare und zugängliche Informationen) und fähig (Qualifikation), die Tätigkeit notfalls selbst durchzuführen?
- Können Sie die KI-Anwendung notfalls ausschalten?

Transparenz

- Ist das Verhalten der Anwendung transparent?
- **Erklärbarkeit:** Ist für Sie nachvollziehbar, wie eine Vorhersage zustande gekommen ist?
- **Interpretierbarkeit (des Modells bei maschinellem Lernen):** Ist das Lernverfahren als Ganzes für Sie transparent?
- **Nachverfolgbarkeit und Reproduzierbarkeit:** Ist sichergestellt, dass das Vorgehen und die Entscheidungen der KI-Anwendung dokumentiert werden?
- **Verlässlichkeit:** Können Sie sich auf die KI-Anwendung verlassen?
- **Korrektheit:** Sind die Ausgaben der KI-Anwendung korrekt?
- **Modellunsicherheit:** Gibt Ihnen die KI-Anwendung Rückmeldung, mit welcher Wahrscheinlichkeit die Ausgaben korrekt sind?
- **Robustheit:** Fängt die KI-Anwendung gestörte oder manipulierte Eingaben ab? Reagiert sie bei unerwarteten Situationen bzw. an den Grenzen des ML-Modells noch korrekt? Fängt sie mögliche Fehler ab (z. B. über „Sanity Checks“)?

Sicherheit

- Ist die Anwendung sicher?
- **Funktionale Sicherheit („Safety“):** Ist sichergestellt, dass die KI-Anwendung die Außenwelt nicht in Gefahr bringt oder schädigt?
- **IT-Sicherheit („Security“)**
  - **Integrität:** Ist die KI-Anwendung vor ihrer Umgebung (inkl. gezielte Manipulationen) geschützt?

– **Verfügbarkeit:** Ist die Verfügbarkeit der KI-Anwendung gewährleistet?

Datenschutz

- Wird der Datenschutz gewahrt?
- Werden die Datenschutz-Grundverordnung (DSGVO) und das Bundesdatenschutzgesetz (BDSG) eingehalten?
- Wurden die folgenden Punkte eingehalten: Einwilligung der Betroffenen, Weiterverarbeitung nur mit Zustimmung, keine unberechtigten Zugriffsmöglichkeiten, weitreichendes und jederzeitiges Widerspruchsrecht, Information über Zweck und Einsatz der personenbezogenen bzw. daraus abgeleiteten Daten, Datensparsamkeit sowie zweckgebundenen Verwendung?

7.8 Übungsfragen: Mensch-KI-System

Zur Überprüfung Ihres Wissensstandes können Sie die folgenden Fragen beantworten.

1. Was ist mit MABA-MABA (oder HABA-MABA) gemeint?
2. Warum kann man eine KI-Anwendung nicht einfach als Kollegin sehen?
3. Welche Metapher von Steve Jobs ist hier vielleicht hilfreicher?
4. Welche Sichtweise haben Sheridan und Verplank (1978) bezüglich der Automation von Tätigkeiten?
5. Schauen Sie sich einmal Ihre Arbeitstätigkeiten an und überlegen Sie, welchen Grad der Automation von Sheridan und Verplank (1978) Sie für die entsprechende Tätigkeit oder Teilaufgabe dieser Tätigkeit akzeptieren würden.
6. Was ist mit Human Control/in/on/out of the Loop gemeint?
7. Schauen Sie sich einmal Ihre Arbeitstätigkeiten an und überlegen Sie, welchen Grad der Automation von Human Control/in/on/out of the Loop Sie für die entsprechende Tätigkeit oder Teilaufgabe dieser Tätigkeit akzeptieren würden.
8. Das Human-Centered Artificial Intelligence (HCAI) von Shneiderman (2020a) unterscheidet sich in einem Hauptpunkt von Sheridan und Verplank (1978) – welchem?

9. Schauen Sie sich einmal Ihre Arbeitstätigkeiten an und überlegen Sie, welche Quadranten (Computerkontrolle, menschliches Können, verlässliche, sichere, und vertrauenswürdige KI) sich für die entsprechende Tätigkeit oder Teilaufgabe dieser Tätigkeit eignen würden.
10. Welche Vorbedingung muss für gute Zusammenarbeit zwischen Mensch und KI gegeben sein?
11. Welche Punkte müssen (u. a. von Poretschkin et al., 2021) bei der Gewährleistung von Autonomie und Kontrolle beachtet werden?
12. Welche Punkte müssen (nach Poretschkin et al., 2021) bei Transparenz/Nachvollziehbarkeit sichergestellt werden?
13. Was umfasst (nach Poretschkin et al., 2021) die Verlässlichkeit eines KI-Systems?
14. Was fällt (nach Poretschkin et al., 2021) unter Sicherheit?
15. Welche Phasen der Interaktion unterscheiden die Richtlinien von Microsoft (Amershi et al., 2019) und welche Punkte sollten jeweils eingehalten werden?
16. Was sind die „Prometheus Prinzipien“ von Shneiderman (2020a)?
17. Wenn Sie eine KI-Anwendung verwenden, dann schauen Sie einmal, wie gut diese Kriterien (die von Amershi et al., 2019 und die von Shneiderman, 2020a) eingehalten wurden.

7.9 Aufgaben zum eigenen Anwendungsfall

In diesem Aufgabenteil schauen Sie sich die Interaktion zwischen dem Nutzer und der KI-Anwendung an.

- Bewerten Sie die Anwendung mit den Kriterien von Amershi et al. (2019) sowie den zusätzlichen Kriterien von Shneiderman (2020a).
- Berücksichtigen Sie hier auch das Kriterium „Sicherheit“ (KI-Anforderungen). Bewerten Sie die Anwendung. Falls hier Defizite vorliegen – welche Ideen haben Sie, wie man die Anwendung bezüglich funktionaler und IT-Sicherheit (letzteres sowohl was Integrität und Verfügbarkeit betrifft) verbessern könnte?
- Modifizieren Sie ggfs. Ihre Anwendung, sodass die Mensch-KI-Interaktion besser funktionieren kann.

Literatur

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-ai interaction. In *Proceedings of the 2019 CHI conference on human factors in computing systems* (S. 1–13). Association for Computing Machinery. https://doi.org/10.1145/3290605.3300233.

Apple Inc. (2022). Human Interface Guidelines. *Apple Computer, Inc.* https://developer.apple.com/design/human-interface-guidelines/guidelines/overview/. Zugegriffen: 15. Okt. 2022.

Fitts, P. M. (1951). *Human Engineering for an effective air-navigation and traffic-control system.* Ohio State University Research Foundation.

Gode, A., & Franke, T. (2019). KI in der ÖV – Der Computer in Erklärungsnot? In *Tagungsband der Veranstaltung am 20. März 2019 Künstliche Intelligenz – Politische Ansätze für eine moderne Gesellschaft* (S. 21–22). opencampus.sh. http://resources.opencampus.sh/190320_KI-Tagungsband.pdf. Zugegriffen: 15. Okt. 2022.

IBM. (2019). *Everyday Ethics for Artificial Intelligence.* IBM Corp. https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf. Zugegriffen: 15. Okt. 2022.

Microsoft. (2019). Guidelines for human-AI interaction. *Microsoft*. https://www.microsoft.com/en-us/research/project/guidelines-for-human-ai-interaction/. Zugegriffen: 15. Okt. 2022.

Poretschkin, M., Schmitz, A., Akila, M., Adilova, L., Becker, D., Cremers, A., Hecker, D., Houben, S., Mock, M., Rosenzweig, J., Sicking, J., Schulz, E., Voss, A., & Wrobel, S. (2021). *Leitfaden zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz–KI-Prüfkatalog.* Fraunhofer IAIS. www.iais.fraunhofer.de/ki-pruefkatalog. Zugegriffen: 15. Okt. 2022.

Rohde, N. (25. October 2017). In Australien prüft eine Software die Sozialbezüge – und erfindet Schulden für 20.000 Menschen. *Algorithmenethik*. https://algorithmenethik.de/2017/10/25/in-australien-prueft-eine-software-die-sozialbezuege-und-erfindet-schulden-fuer-20-000-menschen/. Zugegriffen: 15. Okt. 2022.

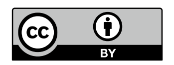
Sheridan, T. B., & Verplank, W. L. (1978). *Human and Computer Control of Undersea Teleoperators.* Massachusetts Institute of Technology.

Shneiderman, B. (2020a). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human-Computer Interaction, 36*(6), 495–504. https://doi.org/10.1080/10447318.2020.1741118.

Shneiderman, B. (2020b). Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy human-centered AI systems. *ACM Transactions on Interactive Intelligent Systems, 10*(4). https://doi.org/10.1145/3419764.

Shneiderman, B. (2021). Responsible AI: Bridging from ethics to practice. *Communications of the ACM, 64*(8), 32–35. https://doi.org/10.1145/3445973.

Voigt, K.-I. (2018). Automatisierung. *Gabler Wirtschaftslexikon.* https://wirtschaftslexikon.gabler.de/definition/automatisierung-27138/version-250801. Zugegriffen: 15. Okt. 2022.


BY



Erklärbare KI

8

Zusammenfassung

Mit der zunehmenden Komplexität von KI-Systemen steigt auch die Notwendigkeit, die Funktionalität der genutzten Systeme verständlich zu machen. Aufgabe dieses Kapitels ist es diesen Aspekt näher zu beleuchten. Zentral ist dabei die Frage wie KI-Systeme so gestaltet werden können, dass sie nicht nur ihre intendierte Funktion erfüllen, sondern die Nutzenden so über ihre Funktionsweise informieren, dass diese die Systeme verstehen, nutzen und steuern können. Zentrale Inhalte sind dabei die Definition von Erklärungen, verschiedene Arten von Erklärungen, sowie Strategien für die bestmögliche Implementierung.

8.1 Einleitung

KI-Systeme sind inzwischen in der Lage, komplexe Aufgaben zu bearbeiten. So komplex, dass die menschlichen Nutzenden manchmal nur verwundert danebenstehen und sich fragen: „Wie hat das System das geschafft?“ Zumindest, wenn alles geklappt hat. Denn nicht immer arbeiten KI-Systeme genau so, wie man es erwartet hat. Vielleicht wurde eine Spam-E-Mail fälschlicherweise als dringlich gekennzeichnet, eine Transaktion wurde als verdächtig gekennzeichnet – aus unbekannten Gründen. Schnell wird die Frage „Wie ist das System vorgegangen?“ wichtig und grundlegender: „Warum hat das System eine bestimmte Entscheidung getroffen?“

Systeme, die in der Lage sind, die Klärung dieser Frage zu unterstützen, fallen – grob gesagt – in die Kategorie der Erklärbaren KI. Erklärbar bedeutet, dass es eben möglich ist, Begründungen für spezifische Entscheidungen oder

M. Heine et al., *Künstliche Intelligenz in öffentlichen Verwaltungen*, Edition eGov-Campus, https://doi.org/10.1007/978-3-658-40101-6_8

das generelle Vorgehen eines Systems zu erhalten. Manche Systeme sind leicht erklärbar – zum Beispiel, wenn die Entscheidungen eines Systems immer auf spezifischen Regeln beruhen. Dann kann das System darauf verweisen, welche Regeln es angewendet hat.

Herausforderung KI
Das ist aber nicht immer möglich – gerade neuere Technologien, z. B. neuronale Netze arbeiten nicht mit expliziten Regeln, die dem typischen Nutzenden als Erklärung ausreichen würden. Daher ist die Entwicklung zusätzlicher Technologie – Erklärbarer KI – notwendig. Denn manchmal können Erklärungen oder zumindest Erklärbarkeit die Voraussetzung dafür sein, dass Systeme eingesetzt werden. Ein System erklärbar zu machen, kann Auswirkungen auf die Vorgehensweise, Leistungsfähigkeit und Anwendbarkeit des Systems haben (vgl. Adadi & Berrada, 2018) – deswegen ist es wichtig, sich frühzeitig mit diesem Thema zu beschäftigen.

Im folgenden Kapitel wird das Konzept der Erklärbarkeit eingeführt und durch welche Technologien diese hergestellt werden kann.

8.2 Fallbeispiele

8.2.1 SchreibFix – Beispiel 1

Ludwig bearbeitet eine E-Mail im Bereich der Finanzaufsicht. Er wollte einen Vorgang an eine Kollegin weiterleiten, die ein ähnliches Problem wie er hat und sich einmal seinen Fall anschauen wollte. Als er SchreibFix anweist, eine entsprechende E-Mail vorzubereiten, ist er verwundert über das Ergebnis: „Ungefähr 2500 Worte wurden aus Datenschutzgründen geschwärzt."

Irritiert scrollt Ludwig durch die E-Mail: und in der Tat – in fast allen Anhängen wurden Schwärzungen vorgenommen. Viele davon kann er gut verstehen – Adressen, Namen und andere, personenbezogene Daten – aber in einigen Passagen wurden auch seine eigenen Analysen geschwärzt. Er fragt seine Kollegin Ingrid, ob sie wisse, weshalb hier Schwärzungen vorgenommen worden seien. Sie verneint: „Am Ende weiß SchreibFix das aber sicherlich besser als wir beide. Ich würde es einfach so losschicken."

Ludwig gibt sich damit nicht zufrieden und schaut sich den Text nochmal an. Es wird kein Tooltip für die Schwärzungen angezeigt und auch sonst gibt es keine Informationen – außer der groben Anzahl an geschwärzten Worten. Er ruft bei der Entwickler-Firma von SchreibFix an und fragt nach. Das System, so erfährt

er, würde jedes Wort mit einem Datenschutzscore gewichten und dann ab einer gewissen Grenze schwärzen. Allerdings kann man ihm nicht sagen, wie genau dieser Score berechnet würde, weil das „für jede Behörde und teilweise Abteilung individuell trainiert wird." Allerdings erhält er Tipps, um die Schwellenwerte für das Schwärzen anzupassen.

Die nächsten zwei Stunden verbringt Ludwig damit, an den Schwellenwerten herumzuspielen. Am Schluss ist er ganz zufrieden mit dem Ergebnis: nur die personenbezogenen Daten sind weg, der Rest ist erhalten geblieben. Er speichert sich diese Einstellung ab und hofft, dass das beim nächsten Mal noch genauso gut funktioniert.

8.2.2 Memoriali – Beispiel 2

Johanna überlegt nicht lange, dann schreibt sie: „…mit diesen Spezifikationen nicht akzeptieren. Bitte verbessern Sie Ihren Antrag und reichen Sie ihn nach Prüfung erneut ein." Das ist der Stand der Dinge, zumindest sagt das Memoriali. Sie setzt sich an die Bearbeitung eines anderen Falles und freut sich, dass der Prozess inzwischen so viel schneller über die Bühne geht. Wenige Stunden später klingelt das Telefon: „Entschuldigen Sie mal, aber was ist das denn für eine Rückmeldung? Bitte versuchen Sie es noch mal? Wo sind wir denn hier, beim Antragslotto? Können Sie mir bitte mal sagen, was an meinen Dokumenten nicht stimmt?"

Johanna öffnet etwas panisch Memoriali und sucht den Antrag heraus, den sie zuvor bearbeitet hatte. Da steht es klar: „Memoriali empfiehlt eine Ablehnung des Antrages, da die eingereichten Daten unvollständig sind/keine hinreichende Qualität besitzen." Das wird schwer zu vermitteln sein – vielleicht kann sie die Daten ja schnell selbst durchklicken. „Klar, bitte gedulden Sie sich nur einen Augenblick…" 141 Seiten PDF, darunter sogar Vergleichsangebote verschiedener Baufirmen. Mehrere Sachverständigenberichte, das sieht doch gut aus. „Hören Sie? Mein System ist gerade sehr langsam, aber ich rufe Sie heute noch zurück." Die Fehlersuche beginnt.

Ihre Kollegin Nadja kann ihr helfen – sie hatte schon mal ein ähnliches Problem. „Es gibt unter ‚Ergebnis', dann ‚Details' und dann ‚Detailbericht anzeigen' alle Fehlermeldungen und Erklärungen zu dem Fall." Johanna ist erleichtert – und in der Tat findet sie schnell das Problem: ein Besitznachweis über das baulich zu verändernde Gebäude fehlte einfach – alle andren Unterlagen waren tadellos.

Das hätte das Programm auch direkt sagen können, ärgert sie sich. Einen unangenehmen Anruf später ist das Problem behoben, die Urkunde trifft nach kürzester Zeit ein und Memoriali hat nichts mehr zu beanstanden – was eine Aufregung.

8.3 Warum erklären?

In den beiden Fallbeispielen wird deutlich, dass es unter Umständen schwierig ist, zu verstehen, wie ein intelligentes System zu seinem Ergebnis gekommen ist. Das kann erhebliche Auswirkungen darauf haben, ob und auch wie ein intelligentes System zum Einsatz kommt. Fehlende Nachvollziehbarkeit, fehlendes Vertrauen und ähnliche, psychologische Faktoren in der Zusammenarbeit mit intelligenten Systemen können sogar dafür sorgen, dass die Performance niedriger ist als vorher. Erklärungen können jedoch Abhilfe schaffen.

8.3.1 Wie Erklärungen helfen können

Erklärungen können auf viele verschiedene Arten dabei helfen, intelligente Systeme optimal am Arbeitsplatz (und in anderen Bereichen) zu integrieren (vgl. dazu Rutjes et al., 2019). Im Folgenden sind einige Ansätze mitsamt Fallbeispielen aufgelistet:

Vertrauensbildung
In dem Moment, in dem wir Teile unserer Arbeit an ein intelligentes System abgeben, machen wir uns von der Qualität dessen Arbeit abhängig. Das bedeutet, dass das Ergebnis unserer eigenen Arbeit davon abhängig ist, wie gut das System arbeitet. Dies ist eine Form von Vulnerabilität, schlechte Systeme könnten unsere Leistung schädigen oder schmälern. In Situationen wie diesen ist es wichtig, dass wir Vertrauen in unser Gegenüber – in diesem Fall das System – haben. Damit ist zum Beispiel das Vertrauen in die Leistungsfähigkeit des Systems gemeint. Kann das System überhaupt die Aufgaben übernehmen, die ich ihm geben will? Weiß das System über alle relevanten Informationen Bescheid und kann diese berücksichtigen? Damit Vertrauen gebildet werden kann, kann es unter Umständen wichtig sein, dass die Arbeitsweise eines Systems klar wird. Nicht immer reicht es aus, zu sehen, dass „das System funktioniert“, also reliabel ist. Besonders in Fällen, bei denen ein Fehler zu ernsthaften Konsequenzen führen kann, ist es notwendig, die genaue Funktionsweise des Systems zu erörtern. Daher können Erklärungen positiv auf das Vertrauen, das einem System entgegengebracht wird, wirken.

Fairness sicherstellen & Regeln einhalten
Nicht immer ist es Ziel eines Systems, alle verfügbaren Daten auch in ein Ergebnis zu integrieren. Nehmen wir zum Beispiel ein System, welches im Bewerbungsverfahren unterstützen soll. Dieses System kann vorhergehende Arbeitsplätze, Universitätsabschlüsse, Zeugnisse oder auch die metrischen Ergebnisse von Arbeitsproben mit in seine Entscheidung einbeziehen. Es gibt allerdings verschiedene Faktoren, die das System nicht in die Entscheidung mit einbeziehen darf – das hat zum Beispiel rechtliche Gründe. Nehmen wir als Beispiel hier die Wahrscheinlichkeit dafür, dass eine Person schwanger ist oder den Grad der Schwerbehinderung, den eine Person besitzt. Ein System, das nicht darlegt, ob und wie es diese Informationen verarbeitet, kann beziehungsweise darf mitunter gar nicht eingesetzt werden. Auch hier kann die Erklärung – zum Beispiel, dass diese Daten nicht für eine Berechnung genutzt werden – unterstützend wirken und Fairness sicherstellen.

Kooperationsfähigkeit verbessern
Auch heutzutage noch sind viele intelligente Systeme auf die Kooperation mit Menschen angewiesen. Das kann zum Beispiel daran liegen, dass die Menschen dafür verantwortlich sind, welche Informationen das System zur Verfügung hat. Oder aber das System gibt ein Ergebnis aus, mit dem der Mensch danach weiterarbeiten muss. In jedem Fall kann die Kooperation zwischen Mensch und KI-System davon profitieren, dass Erklärungen über die Wirkweise des Systems zur Verfügung stehen. Nehmen wir als Beispiel einen Antrag in Memoriali, bei dem das System am Schluss eine große Unsicherheit ausgibt. Das bedeutet, dass das System sagt: „Ich habe eine Vermutung, bin mir aber zu unsicher, um eine klare Empfehlung auszusprechen.“ In diesem Fall wäre es von Vorteil, wenn das System gleichzeitig erklären könnte, auf welche Informationen es zurückgreift. Dann wäre auch die Information enthalten, welche dieser Daten unter Umständen fehlen oder die Sicherheit des Systems in seiner Einschätzung senken. Sachbearbeitende könnten dann gezielt versuchen, diese Informationen zu verbessern.

Eigene Fähigkeiten erwerben & aufrechterhalten
Es kann auch vorkommen, dass das System Aufgaben übernimmt, für die eigentlich der Mensch ausgebildet worden ist. Es könnte das Ziel sein, dass der Mensch auch nach wie vor in der Lage ist, diese Aufgaben selber auszuführen – beispielsweise, weil sie auch einmal zeitkritisch sein könnten. Ein Beispiel wäre die Gestaltung oder der Aufbau von Durchsuchungsbefehlen – sollte das System einmal aus einem unvorhersehbaren Grund nicht verfügbar sein, kann nicht auf eine Reparatur gewartet werden. Erklärungen, die Einblick in die Arbeit des Systems geben, können dafür sorgen, dass es auch in Zukunft den Sachbearbeitenden leichter fällt, die Aufgaben

des Systems auch alleine zu bewältigen. Das kann man sich in etwa so vorstellen, wie es auch in menschlicher Zusammenarbeit ist. Wenn Sie eine erfahrene Person haben, die nicht darüber spricht, wie sie arbeitet, werden die Personen, die mit ihr arbeiten, diese Arbeit auch nicht übernehmen können. Gibt diese Person allerdings explizit Erklärungen zu ihrer Arbeitsweise ab und kann vielleicht sogar spezifische Nachfragen beantworten, besteht eher die Möglichkeit, dass es zu einem Fähigkeitserwerb kommt. Erklärungen können also auch dazu dienen, dies in Teams zu gewährleisten, in denen Menschen und intelligente Systeme zusammenarbeiten.

Wissenserwerb
Gerade weil intelligente Systeme teilweise in der Lage sind, Aufgaben zu lösen, die von Menschen nicht so leicht bewerkstelligt werden können, kann es spannend sein, wenn sie ihr Vorgehen erklären können. Stellen wir uns einmal vor, es wäre nur mithilfe eines neuronalen Netzes möglich, das Verkehrsaufkommen an einer bestimmten Straße präzise vorherzusagen. Die Art und Weise, wie dieses System das macht, könnte anderen Kommunen dabei helfen, auch bei sich und unter anderen Gegebenheiten solche Vorhersagen zu treffen. Ein eigenes Netz zu trainieren ist vielleicht wegen knapper oder unvollständiger Daten nicht möglich.

8.3.2 Was ist eigentlich eine Erklärung – und was nicht?

Grundsätzlich stellt jede Erklärung eine Form von Information dar, die zwischen zwei Partnern kommuniziert wird. Ob eine kommunizierte Information eine Erklärung ist oder nicht, hängt dabei davon ab, in welcher Form sie kommuniziert wird und in welchem Zustand diese Information zuvor beim Gegenüber vorhanden war. Wird eine Information z. B. „Der Boden ist nass." präsentiert, um eine Schlussfolgerung wie z. B. „Es regnet." zu begründen, könnte nach Toulmin (2003) von einer Erklärung gesprochen werden. Die Offenlegung einer Prämisse, von Fakten oder Ursachen für Ereignisse oder Schlussfolgerungen sind also z. B. Erklärungen.

Es könnte allerdings auch sein, dass diese Information dem Gegenüber bereits vorliegt und die Frage einer Erklärung sich nicht auf die genutzten Informationen, sondern auf die angewendeten Regeln, die zur Schlussfolgerung führen, beziehen. In diesem Beispiel würde die Offenlegung der nach Toulmin als Schlussregel oder warrant bezeichneten Aussage: „Dadurch, dass bei Regen Wasser auf den Boden fällt, wird dieser nass." als Erklärung dienen.

Eine Erklärung hängt also immer davon ab, welcher Teil einer z. B. Argumentationsstruktur vom Gesprächspartner oder dem Nutzenden angefordert wird.

Auch die Art und Weise der Erklärung, die im kommenden Kapitel beispielhaft besprochen werden soll, hat einen Einfluss. Für manche Schlussfolgerungen existieren z. B. so viele zugrunde liegenden Informationen oder Schlussregeln, dass nicht alle für eine Erklärung genannt werden können – oder dies schlichtweg unmöglich zu bewältigen wäre. Wenn jemand z. B. fragt: „Wieso wurde dieser Antrag angenommen?“ müssten sämtliche Daten offengelegt werden. Die Frage könnte hier auch anders gestellt werden, z. B. „Welche Information war entscheidend dafür, dass der Antrag angenommen wurde?“ oder „Was hätte passieren müssen, damit der Antrag abgelehnt worden wäre?“ Damit richtet sich die Forderung einer Erklärung auf ein bestimmtes Set an Informationen oder Schlussregeln.

8.3.3 Unterschiedliche Level von Erklärungen

Im Bereich der Mensch-KI-Interaktion gibt es unterschiedliche Levels, auf denen solche Systeme erklärt werden können und erklärbar sein müssen. Für unterschiedliche Nutzergruppen werden unterschiedliche Levels benötigt:

- Auf der globalen Ebene helfen Erklärungen dabei, zu verstehen, wie das System generell arbeitet. Dabei geht es darum, wie die Technologie aufgebaut ist. Im Falle des maschinellen Lernens könnte z. B. der Ansatz eines neuronalen Netzes erklärt werden. Auch die Information, aus wie vielen neuronalen Schichten ein bestimmtes Netz besteht oder mit welchen Daten ein Modell trainiert wurde, sind globale Erklärungen.
- Im Beispiel SchreibFix wurde das System über ein neuronales Netz trainiert. Als globale Erklärungen könnte hier also dargestellt werden, wie das Netz aufgebaut ist, wie genau es zu den Daten passt. Im Abschnitt Metadaten von Ergebnissen wurde bereits erklärt, dass Werte wie der F1-Wert oder die Accuracy dazu genutzt werden können, die Eignung eines Modells für bestimmte Daten zu erkennen. Insbesondere der Vergleich der Accuracy für z. B. unterschiedliche Datensätze kann Aufschluss darüber geben, welche Bereiche Modelle besonders gut oder eben nicht hinreichend abdecken können.
- Die globale Erklärungsfähigkeit eines Modells spielt insbesondere in der Entwicklung des Modells und vor der Nutzung des Modells eine zentrale Rolle. Globale Bewertungen der Performance können dabei helfen, Verbesserungen

an Machine-Learning-Modellen vorzunehmen und so entsprechende Parameter zu konfigurieren. Sie können auch aufzeigen, in welchen Bereichen ein Modell Schwächen hat und somit zeigen, wo Daten fehlen, um ein möglichst umfangreiches Training zu gewährleisten. Wenn z. B. verschiedene Modelle und Ansätze zur Auswahl stehen, sind Methoden der globalen Erklärfähigkeit hilfreich, um eine Entscheidung zu ermöglichen.
- Globale Erklärungen sind daran zu erkennen, dass sie keinen Mehrwert haben, um eine spezifische Entscheidung von einer anderen abzugrenzen.

8.3.4 Übung

1. Welche dieser folgenden Punkte können durch die Erklärbarkeit von KI-Systemen erreicht werden ?
   a) Herstellung fairer Bedingungen
   b) Unterstützung der Vertrauensbildung
   c) Erhöhung der Präzision des Systems
   d) Verringerung des Implementierungsaufwands
2. Welche Voraussetzung erhöht die Kooperationsfähigkeit eines Systems?
   a) Dem Menschen ist klar, welche Daten es verarbeitet hat
   b) Der Mensch weiß, wer das System programmiert hat
   c) Das System zeigt keine Unsicherheiten in der Ausgabe an
   d) Das System zeigt immer numerische Ergebnisse, keine Kategorien
3. Welche Aussage zur Erklärung „Dieser Antrag wird abgelehnt, da er nicht fristgerecht einging“ ist korrekt?
   a) Diese Erklärung wirkt sich negativ auf die Vertrauensbildung aus, da sie negativ formuliert ist.
   b) Diese Erklärung ist lokal, da sie sich auf einen bestimmten Antrag bezieht
   c) Diese Erklärung ist global, da die Frist eine allgemeine Variable im System ist
   d) Keine der vorherigen Aussagen ist korrekt

8.4 Wie erklären?

In den beiden Fallbeispielen wird deutlich, dass es unter Umständen schwierig ist, zu verstehen, was und wie ein intelligentes System gearbeitet hat. Das kann erhebliche Auswirkungen darauf haben, ob und auch wie gut ein intelligentes System arbeitet, wenn es im Einsatz ist. Fehlende Nachvollziehbarkeit, fehlendes Vertrauen und ähnliche, psychologische Faktoren in der Zusammenarbeit mit intelligenten Systemen können sogar dafür sorgen, dass die Performance niedriger ist als vorher. Erklärungen können jedoch Abhilfe schaffen.

8.4.1 Methoden der Erklärbarkeit

Es gibt verschiedene Methoden, wie Ergebnisse eines intelligenten Systems erklärt werden können (vgl. zum Beispiel Nushi et al., 2018; Ribeiro et al., 2016; Weld & Bansal, 2019). Das hängt unter anderem auch davon ab, welche Form des Outputs das System konkret erzeugt und mit welchen Inputdaten es zuvor gearbeitet hat. Werden zum Beispiel Bilddaten verarbeitet, können andere Erklärungsmethoden sinnvoll sein als bei der Verarbeitung von Textdaten oder gar tabellarischen Daten. Bei Bilddaten könnte man zum Beispiel das Bild selbst verändern, um eine Erklärung auch visuell zu repräsentieren. Bei tabellarischen Daten können bestimmte Teile der Tabelle hervorgehoben werden.

Weiterhin unterscheiden sich verschiedene Ansätze, die im Bereich der Erklärbarkeit genutzt werden, in ihrer Interaktivität. So kann eine Erklärung statisch sein und direkt mit einem bestimmten Ergebnis mitgeliefert werden. Sie kann allerdings auch interaktiv gestaltet sein und auf Eingaben der Nutzenden reagieren. Haben die Nutzenden zum Beispiel die Möglichkeit, optionale Ergebnisse zu betrachten und können selber auswählen, welche dieser Optionen sie sich anschauen wollen, dann ist eine Form von Interaktivität gegeben. Dies kann Vor- und Nachteile mit sich bringen.

Die spannendste Frage im Hinblick auf Erklärbarkeit ist jedoch, ob diese auch gegeben ist, wenn keine expliziten Erklärungen gegeben werden. Das könnte zum Beispiel der Fall sein, wenn verschiedene Ergebnisse in einem System miteinander einfach verglichen werden können und durch diesen Vergleich indirekt eine Erklärung entsteht. Das System als solches ist nach wie vor erklärbar, es ist allerdings erforderlich, dass die Nutzenden im System explorieren, um die Information zu erhalten, die sie als Erklärung benötigen. Dies ist zum Beispiel gegeben, wenn Nutzende verschiedene Hotelpreise miteinander vergleichen wollen. Sie haben dort zum Beispiel einfach die Möglichkeit, den Startzeitpunkt

ihrer Reise zu verändern und zu betrachten, wie sich das auf den Gesamtpreis auswirkt. Es gibt keine explizite Erklärung zur Preisberechnung, sie können aber durch Exploration eine für sie ausreichende Erklärung erhalten.

In diesem Kapitel werden exemplarische bekannte Methoden aus dem Bereich der Erklärbarkeit dargestellt.

Darstellung von Trainingsdaten
Auch das objektive Darstellen der Trainingsdaten kann dafür sorgen, dass Nutzende einen Algorithmus besser verstehen. Wenn klar ist, mit welchen Daten ein Netz trainiert worden ist, können z. B. Limitationen abgeleitet werden: Man stelle sich einen Algorithmus vor, welcher das Alter einer Person anhand eines Bildes bestimmen kann, jedoch wurde der Algorithmus ohne Bilder einer bestimmten ethnischen Gruppe trainiert. Eine naheliegende Schlussfolgerung wäre, dass der Algorithmus keine zuverlässigen Ergebnisse für diese Personengruppe liefert.

Darstellung ähnlicher Daten
Insbesondere bei Klassifizierungsaufgaben kann es enorm hilfreich sein, wenn Inputdaten verglichen werden können, die in gewisser Weise ähnlich sind, jedoch zu einer unterschiedlichen Klassifizierung führen. Stellt man die Randfälle im Entscheidungsprozess des KI-Modells dar, so kann man erklären, welche Unterschiede genau für ein anderes Ergebnis ausschlaggebend sind.

Darstellung von Extremfällen
Eine weitere Art der Randfälle für KI-Modelle sind die Extremfälle in Datensätzen. Diese Daten können durch statistische Verteilungen ermittelt und den Nutzenden dargestellt werden. Durch die Darstellung extrem abweichender Daten können Schlussfolgerungen bzgl. der Limitationen des Modells abgeleitet werden: Trainingsdaten die extrem selten oder in Kombination extrem selten vorkommen deuten auf Schwachstellen des Modells hin.

Attribution
Bei der Analyse der Attribution wird überprüft, welchen Einfluss bestimmte Daten, die Teil des Inputs sind, auf das Ergebnis einer KI-Berechnung haben. Ein gutes Beispiel hierfür ist die Analyse von Bildern als Stimuli. Jedes Bild besteht aus einer bestimmten Anzahl an Pixeln. Jedes Pixel wiederum besitzt einen bestimmten Farbcode. Das bedeutet, dass ein Bild das zum Beispiel 512×512 Pixel groß ist, insgesamt rund 260.000 einzelne Pixel mit einer Farbinformation besitzt. Die Frage ist nun: wie groß ist der Einfluss eines einzelnen Pixels oder von einer Gruppe von Pixeln auf das Ergebnis der KI?

Modelldestillation

Die Modelldestillation bezieht sich auf eine Klasse von Erklärungsmethoden, bei denen das in einem trainierten Modell kodierte Wissen in eine Darstellung destilliert wird, die zugänglich für Nutzende ist. Diese Darstellung kann die Form von besser interpretierbaren, maschinellen Lernmethoden annehmen, wie z. B. Entscheidungsbäume. Ein destilliertes Modell lernt im Allgemeinen, die Aktionen oder Eigenschaften eines 'Black-Box' Models über dieselben Daten zu imitieren.

Intrinsische Methoden

Im Idealfall möchten wir Modelle haben, die Erklärungen für ihre Entscheidungen als Teil der Modellausgabe liefern, oder dass die Erklärung leicht aus der Modelarchitektur abgeleitet werden kann. Mit anderen Worten: Erklärungen sollten dem Prozess der Entwicklung von Modellarchitekturen und dem Training inhärent sein. Nicht alle Modelle sind darauf ausgelegt, Erklärungen zu generieren. Wenn sie es nicht sind, kann es schwer sein mit Post-Hoc-Methoden nachvollziehbare Erklärungen zu generieren. Ein bei der Entwicklung bereits auf Erklärungen angelegtes Modell kann hier Vorteile bieten. Dies liegt daran, dass ein intrinsisches Modell nicht nur in der Lage ist, genaue Ausgaben pro Eingabe zu lernen, sondern auch Ausgaben, die eine Erklärung für das Verhalten des Netzes auszudrücken.

8.4.2 Übung

1. Bei welchem dieser Beispiele handelt es sich nicht um eine Erklärung?
   a) Die Berechnung von Preisen wird bei der Buchung einer Leistung angezeigt
   b) Die drei ähnlichsten Bescheide werden bei der Beurteilung eines Bescheids angezeigt
   c) Die Trainingsdaten einer Bilderkennung werden bei einer Bewertung mit dargestellt
   d) Bei allen obrigen handelt es sich um Erklärungen
2. Ein System soll automatisch das Alter von Fenstern schätzen. Es schlägt bei Fenstern, die vor 1700 eingebaut wurden, regelmäßig fehl. Welche Erklärungsmethode eignet sich hier besonders gut und weshalb?
   a) Attribution, damit man sehen kann, welcher Teil des Fensters Probleme macht
   b) Darstellung von Extremfällen, da diese Fenster ganz anders aussehen als aktuelle

   c) Darstellung der Trainingsdaten, da Bilder aus dieser Zeit fehlen könnten
   d) Keine der obrigen Methoden eignet sich
3. Welches der folgenden Modelle ist nicht intrinsisch erklärbar?
   a) Ein tiefes, neuronales Netz mit weniger als 12 Layern
   b) Ein Baumdiagramm mit mehr als 15 Verzweigungen
   c) Ein Bayes-Netz mit über 30 Knotenpunkten
   d) Keines der genannten Modelle ist erklärbar

8.5 Counterfactual Explanations

In dieser Sektion geht es um eine bestimmte Form von Erklärung, die sich im Bereich des Machine Learnings, erklärbarer KI – aber letztlich auch in unserem Alltag – großer Beliebtheit erfreut: Counterfactual Explanations (vgl. Sokol & Flach, 2018). Was ist das genau?

8.5.1 Beispiel Counterfactual Explanations

„Lara arbeitet gerade an einem komplizierten Fall in Memoriali. Sie ist etwas verunsichert, aber froh, dass das Programm unterstützt, denn sie arbeitet zum ersten Mal in diesem Bereich. Dennoch ist sie etwas verwundert, als das Programm ihr vorschlägt, den Antrag abzulehnen. Sie schaut sich alle Dokumente noch mal an, findet sie aber vollständig und passend. Was könnte diese Entscheidung verursacht haben? Sie nutzt eine neue Funktion: ‚Antragsannahme simulieren'. Dadurch berechnet Memoriali, welche Änderungen in den vorliegenden Daten dafür gesorgt hätten, dass der Antrag angenommen wird. Nach zwei Minuten kann sich Lara den simulierten Antrag anschauen und wird auf die Unterschiede hingewiesen. Insbesondere einen: ‚Datum Antragstellung' – ah, gut, daran lässt sich nun nichts mehr ändern. Sie deaktiviert diesen Punkt für die Simulation und gibt den Antrag noch mal neu in das System. Diesmal erscheint das ‚Datum Baubeginn'. Eine Woche später müsste das Ganze starten, anscheinend, weil die entsprechenden Fristen nicht beachtet worden sind. Damit ist Lara zufrieden – sie lehnt den Antrag nicht ab, sondern akzeptiert ihn mit der Korrektur des Baubeginns."

8.5.2 Was ist Counterfactual Explanation?

Im obigen Beispiel wurde eine Erklärung auf eine besondere Art und Weise gegeben – es wurde geschaut, welche Ausgangsbasis zu einem anderen, definierten Ergebnis geführt hätte. Diese „Was müsste passieren, damit…“-Form der Erklärung bezeichnet man auch als Counterfactual Explanation. Eine Counterfactual Explanation definiert sich über zwei Punkte: 1) sie stellt einen Ausgangspunkt dar, der möglichst nahe an dem liegt, den man gerade untersuchen will und 2) ist ein konkretes Beispiel, d. h. es liegt ein konkreter Wert für z. B. alle Input-Features vor. Der Begriff counterfactual beschreibt genau diese Tatsache: man betrachtet Werte, die „entgegen der vorgefundenen Fakten“ analysiert werden.

Diese Form der Erklärung beruht oft darauf, dass der nächste Datenpunkt, das nächstliegende Beispiel, das zu einem anderen Ergebnis führt als das zu untersuchende, herangezogen wird. Nehmen wir als Beispiel den oben geschilderten Fall aus Memoriali. Hierbei wird der aktuelle Antrag als Ausgangspunkt betrachtet. Dieser erhielt als Ergebnis eine Ablehnung. Um eine Counterfactual Explanation zu bieten, sucht das System unter allen anderen Anträgen nach einem, der möglichst nahe zu dem vorliegenden ist – es kann diesen tatsächlich gegeben haben, das System kann diesen aber auch simulieren.

Wurde dieser Punkt gefunden, kann er als Erklärung dargestellt werden. Es kann dabei auch vorkommen, dass es nicht nur eine Counterfactual Explanation gibt, sondern mehrere, weil sich z. B. unterschiedliche Parameter abändern lassen, die jeweils zu einem anderen Ergebnis führen. Es kann auch sein, dass sich mehrere Parameter verändern müssen, damit ein anderes Ergebnis herauskommt. Denken Sie einmal daran, wenn jemand Sie im Winter beim Schneetreiben fragen würde, was anders sein müsste, damit Sie ein T-Shirt tragen würden – vermutlich würde „Höhere Temperatur“ als Änderung nicht ausreichen, sondern es müsste auch „Kein Niederschlag“ gegeben sein.

8.5.3 Vor- und Nachteile von Counterfactual Explanations

Die Nutzung dieser spezifischen Art, Erklärungen zu gestalten, bringt Vor- und Nachteile mit sich. Ein großer Vorteil ist, dass Counterfactual Explanations eine natürliche Art der Frage nachahmen – die Frage: „Was wäre wenn?“ und daher relativ leicht zu verstehen ist, wie diese Erklärungen funktionieren. Das sogenannte „Mentale Modell“ davon, wie ein System arbeitet, basiert auf der Fähigkeit, in seiner Vorstellung z. B. verschiedene Manipulationen oder andere

Parameter auszuprobieren, um zu simulieren, was ein System tun würde. Counterfactual Explanations sind genau darauf ausgerichtet, diesen Prozess in der Interaktion mit einem System zu ermöglichen und helfen so bei einem einfachen Aufbau eines mentalen Modells des KI-Systems.

Außerdem können Counterfactual Explanations auf den Anwendungsfall angepasst werden. Das bedeutet, dass Erklärungen sehr gut von z. B. Entwickelnden gestaltet werden können, indem sie z. B. auf besonders wichtige Features aufmerksam machen und dort die Möglichkeit zur Exploration geben. Ein Beispiel dafür wäre, dass z. B. in SchreibFix der Betreff einer E-Mail eine besondere Rolle spielt. Die Software könnte explizit vorschlagen, bei Erklärungsbedarf den Titel zu ändern oder optionale Titel anzuzeigen und damit optimal das Nutzendenverhalten unterstützen.

Diese konkreten Handlungsmöglichkeiten im Rahmen von Counterfactual Explanations erlauben auch, bestimmte Hypothesen zu untersuchen. Falls z. B. angenommen wird, dass ein Input-Feature wie die Postleitzahl bei der Bewertung eines Antrages relevant wurde, könnte überprüft werden, ob die Änderung einen Einfluss auf das Ergebnis hat. Dadurch, dass fallabhängige Hypothesen gebildet und überprüft werden können, könnte so ermöglicht werden, dass Nutzende aktiv die Vertrauenswürdigkeit eines Systems überprüfen.

Dennoch eignen sich Counterfactual Explanations nicht immer, da sie auch einige Nachteile mitbringen. Hier seien exemplarisch genannt:

- Durch ihre Natur beziehen sich solche Erklärungen nur auf konkrete Beispiele und können damit nicht gut für strukturelle Zusammenhänge genutzt werden. Insofern können nicht alle Hypothesen überprüft werden. Wird z. B. angenommen, dass die Postleitzahl generell einen negativen Einfluss auf Anträge haben kann, müssten sehr viele Fälle verglichen werden. Diese Form der Erklärung wird insbesondere auf Ebene lokaler Erklärung genutzt und nicht als globale Erklärung.
- Die Auswahl der Counterfactual Explanation, die konkret angezeigt wird, ist nicht immer einfach. So ergeben manchmal naheliegende Daten, die zu einer anderen Erklärung führen, keinen Sinn, wie im Beispiel zu sehen ist. Bestimmte Daten können oder sollen ggf. nicht verändert werden. Oder Inputs sind voneinander abhängig, z. B. das Alter eines Fensters und ob es unter den Denkmalschutz fällt oder nicht. Dementsprechend müssen gegebene Erklärungen sorgfältig ausgewählt werden, damit sie nicht nur hypothetisch ein Ergebnis liefern, sondern auch als potenzielle Alternative sinnvoll erscheinen.
- Wie viele Formen der Erklärung bieten besonders Counterfactual Explanations die Möglichkeit, das System zu manipulieren. Wenn das System klar darstellt,

was fehlt, um z. B. bei einem Antrag auf Schwerbehinderung eine gewisse Einstufung zu erhalten, könnte dies dazu führen, dass Anträge explizit so gestaltet werden, um diese Einstufung zu erhalten, ohne dass sich der dahinterliegende Sachverhalt ausreichend ändert. Dieses Thema wird im Bereich KI & Ethik vertiefend diskutiert.

8.5.4 Übung

1. Mit welcher Frage lässt sich „Counterfactual Explanation“ gut beschreiben?
   a) Was wäre wenn … ?
   b) Wann war klar, dass … ?
   c) Wie sicher ist, dass … ?
   d) Woher ist sicher, dass … ?
2. Was ist **kein** Vorteil von Counterfactual Explanations?
   a) Angepasst an natürliche Art, Fragen zu stellen
   b) Untersuchung bestimmter Hypothesen einfach möglich
   c) Leicht in der Programmierung umzusetzen
   d) Einfach an spezifischen Kontext anpassbar
3. Wie könnte eine „Counterfactual Explanation“ dazu beitragen, dass ein System manipuliert wird?
   a) Es ist klar, welche Eingaben verändert werden müssen, um ein anderes Ergebnis zu erreichen
   b) Es werden Informationen über die Trainingsdaten bekannt, wenn Counterfactual Examples dargestellt werden
   c) Bei vielen solcher Anfragen kann es zu einer Überlastung des Systems kommen, da einzelne Counterfactual Explanations rechenintensiv sind
   d) Diese Form der Erklärung ist im Rahmen von Kreditgeschäften entstanden und daher besonders gut für Betrugsversuche geeignet

8.6 Technologien im XAI Bereich

Der folgende Text stellt eine Technologie aus dem Bereich „Erklärbare KI“ beispielhaft dar (vgl. zu weiteren Ansätzen auch Montavon et al., 2018). Als Beispiel wird ein KI-System betrachtet, welches das Bild eines Fensters betrachtet und die dazugehörige Epoche klassifizieren kann.

8.6.1 Pixel für Pixel Relevanz feststellen

Das neue KI-System in Memoriali ist in der Lage, Fenster zu klassifizieren – je nachdem, aus welcher Epoche sie kommen. Da es spannend wäre zu verstehen, wie diese Ergebnisse entstehen, wird ein Ansatz namens „Sensitivitätsanalyse“ ausprobiert. Er gehört zu den in Abschn. „8.4 Wie erklären?“ vorgestellten Methoden aus dem Bereich der „Attribution“. Bei diesem Ansatz wird versucht, die Relevanz jedes einzelnen Pixels für das Ergebnis zu berechnen. Es entsteht also für jeden Pixel eines Bildes ein „Relevanz-Wert“, der widerspiegelt, welchen Einfluss dieser Pixel auf die Klassifikation hatte.

Nimmt man nun alle Relevanz-Werte eines Bildes zusammen, kann man daraus eine sogenannte „Heatmap“ machen. Dabei werden einzelne Pixel in Abhängigkeit von ihrer Relevanz eingefärbt: sehr relevante Pixel sind z. B. rot, irrelevante Pixel werden blass dargestellt. Diese Heatmap kann dann über das zu untersuchende Bild gelegt werden. So kann besser verstanden werden, welche der Pixel eines Bildes für die Klassifikation relevant waren. Auch die Strukturen eines Bildes können so überprüft werden – ob z. B. ein bestimmter Bogen im Fenster für die Klassifikation relevant war oder nicht (vgl. Montavon et al., 2017).

8.6.2 Dekomposition neuronaler Netze

Schaut man sich die zugrunde liegende Technologie genauer an, gibt es eine große Herausforderung – denn den Relevanzwert zu berechnen ist nicht einfach. Das liegt daran, dass die aus Kap. 3 bekannten „Hidden Layers“ dafür sorgen, dass die Beziehung zwischen einem einzelnen Pixel und dem Ergebnis mathematisch auf verschiedene Varianten bestimmt werden kann. Wie gut diese funktionieren, hängt z. B. davon ab, wie viele Schichten ein neuronales Netz besitzt.

Eine Möglichkeit, dieser Komplexität entgegenzutreten, ist, den Relevanzwert nicht über den direkten Zusammenhang zwischen einem Pixel und dem Ergebnis zu berechnen, sondern die dazwischenliegenden Neuronen in die Berechnung miteinzubeziehen. Tatsächlich ist jede Verbindung zwischen Neuronen bzw. den Input-Features durch eine mathematische Funktion gekennzeichnet. Diese einzelnen Funktionen können im Rahmen eines „divide-and-conquer“-Ansatzes voneinander getrennt analysiert werden. Dies erlaubt eine feinere Betrachtung und eine genauere Einschätzung des Relevanzwertes (vgl. Montavon et al., 2017).

8.6.3 Schichtweise rückwärts durch das Netz

Eine Technologie, bei der dieses Vorgehen gewählt wurde, ist die „Layerwise Relevance Propagation", deren Ziel eben die Analyse des Relevanzwertes ist. Dazu „bewegt" sich das System nach einer Prädiktion Schritt für Schritt rückwärts durch die „Hidden Layers" und betrachtet die entsprechenden Aktivierungen der einzelnen Neuronen (vgl. Bach et al., 2015). Folgendes Beispiel verdeutlicht den Ansatz.

Stellen Sie sich vor, 8 Personen stehen alle an einer Linie und erhalten jeweils eine zufällige Anzahl an schweren Steinen. Nun bewegen diese sich in 8 Schritten über einen sehr weichen Boden – die Tiefe des Fußabdrucks hängt davon ab, wie viele Steine die Person trägt. Bei jedem Schritt werden zwischen den Personen auch Steine hin und her gegeben. Am Ende dieser Prozedur legen alle die Steine vor sich ab – je nachdem, wie die Steine am Ende verteilt sind, bringt das Glück oder Pech… aber darum soll es hier nicht gehen.

Die Personen repräsentieren, auf welche Art Informationen (Steine) durch das Netz wandern, miteinander interagieren und am Schluss ein Ergebnis liefern. Zudem wird deutlich, dass es Interdependenzen, also gegenseitige Abhängigkeiten, zwischen den Personen gibt. Um nun zu verstehen, wie viel Einfluss eine bestimmte Person hatte, kann man fragen: „Wie viel Steine hattest du am Anfang?" Das kann zwar ein Indiz dafür sein, wie wichtig diese Person war, aber wie wir wissen, werden zwischendurch auch Steine getauscht. Eventuell ist eine Person mit wenigen Steinen gestartet, hatte wenige Steine am Ende, aber hat zwischenzeitlich viele Steine transportiert. Daher könnten wir die Zeit rückwärts ablaufen lassen und beobachten, wie die Steine von Person zu Person wandern, indem wir schauen, wie tief jeweils die Fußabdrücke waren. Aus diesen Informationen kann die „Relevanz" einer Person abgeleitet werden.

8.6.4 Übung

1. Zu welcher Form der Analyse zählen Verfahren, bei denen die Relevanz einzelner Pixel durch Okklusion Schritt für Schritt untersucht werden?
   a) Kompositionsanalyse
   b) Sensitivitätsanalyse
   c) Transparenzanalyse
   d) Minimale-Partiale-Analyse

2. Wie nennt man das Ergebnis einer Zusammenstellung von farblich kodierten Pixeln eines Bildes, wobei die Farben die Relevanz der einzelnen Pixel darstellen?
   a) Heatmap
   b) Smartie Graphic (Smarg)
   c) Mosaicmap
   d) Rainbow-Visualization
3. Welche Strategie verwendet man, um die Verbindung einzelner Neuronen zu analysieren?
   a) Divide and Conquer
   b) Slice and Dice
   c) Resize and Reshape
   d) Isolate and Integrate

8.7 Erklärungen evaluieren

8.7.1 Was ist eine „gute" Erklärung?

Jede Erklärung, die ein System gibt – egal ob explizit oder implizit –, sollte das Ziel haben, die Mensch-Maschine Interaktion zu verbessern. Es gibt dabei verschiedene Ziele, wie im Kapitel „8.3 Warum erklären?" bereits dargestellt wurde.

Allerdings ist nicht jede Erklärung hilfreich. Wenn neue KI-Systeme implementiert werden, die eine grundlegende Erklärbarkeit enthalten, müssen diese auch analysiert und betrachtet werden, ob die gegebenen Erklärungen auch tatsächlich positiv zur Interaktion beitragen. Der folgende Beispielfall verdeutlicht, was schiefgehen kann:

Das System „SchreibFix" bewertet die Professionalität einer E-Mail. Die wird dem Nutzenden in Form von 5 Stufen angezeigt – zwischen „unprofessionell" und „hochprofessionell". Das Bestreben der Sachbearbeitenden ist es, möglichst professionelle Texte zu verfassen. Nachdem Alex seine E-Mail geschrieben hat, landet er in Kategorie drei. Als Erklärungen werden ihm Worte in seiner E-Mail angezeigt, die besonders positiv oder negativ zu der Bewertung beitragen. Der Satz „Ich hoffe, dieses Schreiben erreicht Sie bei bester Gesundheit." und auch seine Grußformel wurden vom System nicht beachtet. Das irritiert ihn – kann das System dann überhaupt richtig arbeiten? Er ignoriert weitere Markierungen und schickt die E-Mail ab.

In diesem Fall hat die Erklärung des Programms SchreibFix einen negativen Einfluss auf die Interaktion gehabt, weil sie den Nutzer verunsichert hat. Eine Erklärung führt also nicht automatisch zu einer Verbesserung relevanter Faktoren wie der Wahrnehmung von Fairness, Vertrauenswürdigkeit und Leistungsfähigkeit beim Gegenüber. Sie kann auch zu Verunsicherung führen. Daher ist es wichtig, den Einfluss von Erklärungen zu überprüfen.

Was dabei eine „gute“ Erklärung ausmacht, hängt von dem Ziel ab, das man zu erreichen versucht. Eine Erklärung, welche die wahrgenommene Vertrauenswürdigkeit eines Systems verbessert, zeigt zum Beispiel eher an, wie die Daten verarbeitet werden und welches Gewicht bestimmte Features besitzen; während die Fairness mitunter dargestellt werden kann, indem man den Datensatz, der fürs Training genutzt wurde, darstellt. Die Konsequenz: der erste Schritt, wenn man ein erklärbares System entwickelt oder implementieren möchte, ist die Überlegung, wofür die Erklärung gut sein soll. Der zweite Schritt ist die Überprüfung der Effekte.

8.7.2 Methoden zur Evaluation von Erklärungen

Wie wirkt sich eine Erklärung auf die Arbeit, auf die Interaktion zwischen Mensch und Maschine aus? Es gibt unterschiedliche Verfahren, die hier zum Einsatz kommen können. Einige werden im Rahmen dieses Abschnitts vorgestellt – dabei handelt es sich jedoch nicht um eine vollständige Liste.

Objektive Methoden
Die erste Kategorie bilden dabei Methoden, über die objektive Werte erfasst werden. Diese Werte hängen nicht davon ab, wie eine Person die Interaktion erlebt und können daher gut zwischen Personen verglichen werden.

Das (1) erste Beispiel bildet die Messung der Performance. Dabei geht es darum, wie oft ein Mensch-KI-Team zu einem korrekten Ergebnis kommt bzw. wie schnell diese Ergebnisse erreicht werden. Auch die Accuracy oder der F1-Wert, vorgestellt in Abschn. „4.8 Metadaten von Ergebnissen“, können solch eine Performance darstellen. Erklärungen können den Menschen helfen, die Outputs von KI-Systemen besser zu verstehen und dadurch eine höhere Performance ermöglichen.

Im Falle von Memoriali wäre z. B. die Anzahl korrekt durchgeführter Vorgänge ein Maß der Performance. Nehmen wir aber einmal an, die Performance ist nicht so hoch wie gewünscht und es soll herausgefunden werden, woran das liegen kann.

Dann stellt zweitens (2) die Freeze-Probe-Technik aus dem Bereich der „Situation Awareness“ eine gute Möglichkeit dar, um Zwischenschritte besser verstehen

zu können. Dabei wird definiert, welches Wissen eine Person zu verschiedenen Zeitpunkten der Interaktion haben sollte, um optimal entscheiden zu können. In einem solchen Prozess wird sie dann unterbrochen und nach den Informationen gefragt. Dieses Vorgehen wird in verschiedenen Artikeln aus dem Bereich der Situation Awareness geschildert, z. B. zu finden bei de Winter et al. (2019).

Im Fall von Memoriali könnte so z. B. nach der Sichtung von Plänen die Interaktion gestoppt werden und gefragt werden: „Wie viele Bilder wurden eingereicht?“, „Welches Bild war unvollständig?“ oder „Welchen Bereich des Bildes hat das KI-System als irrelevant markiert?“. Diese Fragen können helfen zu verstehen, ob die Interaktion mit dem System so abläuft wie geplant, oder ob Erklärungen z. B. übersehen werden. Dann können sie auch keinen Einfluss auf die Performance haben.

Ebenso kann drittens (3) auch das Verhalten der Personen beobachtet werden. Falls sich Erklärungen z. B. auf Anfrage anzeigen lassen, könnte man herausfinden, ob das Anzeigen der Erklärung einen Einfluss darauf hat, ob der Vorschlag eines Systems angenommen wird. Im Falle von SchreibFix könnte so überprüft werden, ob Personen nach der Anzeige einer E-Mail als „unzureichend professionell“ die E-Mail nochmal korrigieren oder nicht – ob sie ihr Verhalten also verändern, abhängig vom Ergebnis des Systems. Dies bezeichnet man auch als „reliance“ – verlassen sich Personen in ihrem Verhalten auf die Ergebnisse des Systems oder nicht? Hier ist aber Vorsicht geboten: nicht bei jedem System kann dies gemessen werden, da Personen manchmal ohne ein System bestimmte Aufgaben gar nicht durchführen können.

Weiterhin kann ebenso (4) der Aufwand, oft als Workload bezeichnet, einer Person objektiv gemessen werden. Hierzu gibt es unterschiedliche Methoden – z. B. könnte eine Person, die mithilfe eines KI-Systems das Monitoring in einem Bahnhof übernimmt, parallel eine andere Aufgabe bekommen. Wie gut sie diese Aufgabe erledigt, zeigt an, wie viele mentale Ressourcen durch das Monitoring gebunden worden sind – und wie viele für andere frei sind. Aber auch die Zeit, innerhalb derer auf bestimmte Bildschirmbereiche geschaut wird, kann Aufschluss über den Workload geben.

Subjektive Methoden

Die zweite Kategorie bilden subjektive Methoden. Diese Werte werden dadurch erhoben, dass Personen nach ihrem eigenen Erleben gefragt werden. Das kann zum Beispiel nach der Interaktion mit einem System gemacht werden. Durch die unterschiedlichen Interpretationen von z. B. Begriffen in den Fragen oder eigene Vorstellungen, sollte hierbei auf wissenschaftlich validierte Fragebögen gesetzt und

zudem vorsichtig mit einem Vergleich zwischen z. B. zwei Sachbearbeitenden umgegangen werden.

Das im Bereich „Erklärbare KI“ entscheidende (1) Vertrauen bzw. die wahrgenommene Vertrauenswürdigkeit eines Systems, stellen ein gutes Beispiel für eine subjektive Variable dar. Es gibt eine Reihe von Fragebögen, die sich mit diesem Thema beschäftigen und versuchen, Vertrauen zu erheben. Ein Beispiel dafür stellt der Fragebogen von Jian et al. (2000) dar, welcher z. B. im Bereich des autonomen Fahrens häufig zum Einsatz kommt. Teilnehmende haben hier in der Regel mit einem KI-System die Möglichkeit, mehrere Fragen zu z. B. den Intentionen oder der Einstellung des Systems zu beantworten. Andere Fragebögen wie z. B. die FOST-Skala (Trommler et al., 2018) können auch eingesetzt werden.

Ein Beispiel für den Einsatz von Vertrauens-Fragebögen könnte z. B. sein, zu testen, wie die Erklärungen von SchreibFix genutzt werden. So könnten Sachbearbeitende das automatische Generieren von E-Mails bewerten, zunächst ohne und danach mit Erklärungen. Falls die Erklärungen beeinflussen, inwieweit diese Generierung als vertrauenswürdig erlebt wird, sollte dies in den Antworten deutlich werden.

Auch hier bildet der (2) Workload eine Möglichkeit. Mit dem NASA-TLX kann der subjektiv erfahrene Workload erfasst werden (vgl. Grier, 2015). Damit kann geklärt werden, wie anstrengend eine Aufgabe „empfunden“ wird. Hier kann ein Vergleich mit objektiven Maßen des Workloads eine spannende Möglichkeit darstellen, um zu verstehen, wie das System angenehmer gestaltet werden kann. So könnte z. B. bei Memoriali überprüft werden, ob das System nach der Einführung dazu führt, dass die eigene Belastung als geringer eingeschätzt wird.

Die (3) Zufriedenheit mit gegebenen Erklärungen stellt ebenfalls ein subjektives Maß dar, mit dem Erklärungen generell evaluiert werden können. Sie integriert sowohl die Nützlichkeit und Gebrauchstauglichkeit einer Erklärung als auch den angenommenen Effekt einer Erklärung auf die eigene Arbeit. Eine detaillierte Skala dazu wurde z. B. von Hoffman et al. (2018) vorgeschlagen und kann bei Systemen angewendet werden, mit denen die Nutzenden schon etwas vertrauter sind. Die Zufriedenheit mit Erklärungen kann ein Hinweis darauf sein, weswegen ein System seltener genutzt wird, als es sollte und wie es gegebenenfalls verbessert werden könnte.

Insbesondere, wenn Menschen und KI-Systeme interagieren und kooperativer zusammenarbeiten sollen, ist es wichtig, dass die menschlichen Nutzenden verstehen, wie das System arbeitet. Daher spielt die (4) Nachvollziehbarkeit eines Systems – und wie diese erlebt wird – auch eine wichtige Rolle. Dabei geht es darum, ob die Informationsverarbeitung des Systems als transparent erlebt wird und ob die Art und Weise, wie das System zu Ergebnissen kommt, zugänglich erscheint. Je

nachdem, wie die Kooperation zwischen Mensch und Maschine aufgebaut ist, muss eine niedrige Nachvollziehbarkeit nicht zwingend schlecht sein: nicht jeder Prozess muss zwangsweise nachvollzogen werden können. Bei wichtigen Entscheidungen, die z. B. gesetzeskonform oder fair sein müssen, spielt die Nachvollziehbarkeit eine größere Rolle. Die Messung der Nachvollziehbarkeit überschneidet sich teilweise mit der Messung des Vertrauens, es gibt aber auch explizite Skalen wie z. B. SIPA (Schrills et al., 2021).

Bei allen subjektiven Maßen ist es zudem wichtig, Abweichungen in beide Richtungen zu betrachten: ein System, das zu wenig genutzt wird, weil ihm wenig vertraut wird, erreicht die Produktivität und Effizienz nicht, für die es gemacht wurde. Systeme, die zu viel genutzt werden, weil sogenanntes Übervertrauen vorliegt, sind allerdings ebenso ein Problem, da Fehler ggf. nicht erkannt werden und den KI-Systemen zu viel zugetraut wird.

8.7.3 Übung

1. Welche dieser Aussagen kann man bei subjektiven Fragebögen nicht treffen?
   a) Tool A sorgt für deutlich weniger Stress bei den Personen als Tool B
   b) Die Zufriedenheit von Tool B lässt über die Zeit deutlich nach
   c) Die Geschwindigkeit der Aufgabenbearbeitung wird bei Tool B mit der Zeit besser
   d) Mehr als 70 % geben an, dass Tool A deutlich besser zu verstehen sei
2. Aus welchen Gründen ist es sinnvoll, eine Erklärung zu evaluieren?
   a) Erklärungen beeinflussen die Art und Weise, wie das System arbeitet und müssen daher überprüft werden
   b) Erklärungen können auch verwirrend oder unklar sein und sollten daher evaluiert werden
   c) Erklärungen verzögern mitunter die Geschwindigkeit der Arbeit, ohne einen Mehrwert zu bringen
   d) Es sollte vor Einsatz des Tools gewährleistet werden, dass Erklärungen der DIN Norm (6445-92) zu transparenten Systemen entsprechen
3. Was ist keine objektive Messmethode bei der Evaluation von Erklärbarer KI?
   a) Freeze Probe Technik
   b) Verhaltensbeobachtung
   c) Leistungsmessung
   d) Analyse nach Mayring

8.8 Aufgaben zum eigenen Anwendungsfall

In diesem Abschnitt wurde dargestellt, welche Schritte notwendig sind, um die Zusammenarbeit zwischen Mensch und KI zu verbessern. Dafür werden Methoden der Erklärbarkeit eingesetzt. In dieser Aufgabe sollen mögliche Vor- und Nachteile von Erklärung abgewogen werden.

- Beschreiben Sie, weshalb Ihr System von Erklärungen profitieren kann. Welche Verarbeitungsschritte könnten zu Konflikten führen? Wo könnten Erklärungen die Verlässlichkeit von Ergebnissen verbessern?
- Wählen Sie eines der im Kurs vorgestellten Verfahren zur Erklärung von KI-Ergebnissen aus. Begründen Sie Ihre Entscheidung und stellen Sie dar, wie die Erklärung sich auf die Nutzung des KI-Systems auswirken kann. Schildern Sie dazu eine Beispielsituation, in welcher die Erklärung genutzt werden kann.
- Entwickeln Sie einen Plan, um zu prüfen, ob die Erklärung wie erwartet wirkt. Begründen Sie, welche Variablen Sie dabei beachten und wie Sie diese messen wollen. Erörtern Sie, welchen Stellenwert objektive und subjektive Messverfahren dabei haben.

8.9 Zusammenfassung

In diesem Kapitel wurden unterschiedliche Formen von Erklärbarkeit in intelligenten Systemen vorgestellt. Erklärbarkeit ist eine wichtige Eigenschaft von KI-Systemen, um Interaktionsfaktoren wie Vertrauen, Nachvollziehbarkeit und Kooperation zu beeinflussen. Allerdings gibt es eine Vielzahl an Erklärmethoden, die im Rahmen dieses Kapitels vorgestellt worden sind – nicht alle sind immer anwendbar, sondern hängen vom Kontext ab und der Frage, die durch eine Erklärung beantwortet werden soll.

Zunächst wurde dabei die Frage beantwortet „Warum Erklären“ überhaupt wichtig sein kann. Dabei wurde deutlich gemacht, dass Erklärungen zum Informationsaustausch zwischen Menschen und KI beitragen können. Sie sind notwendig, wenn Menschen nicht wissen, wie ein Ergebnis zustande gekommen ist, wie sie es beurteilen können oder ob das System notwendige Regeln eingehalten hat.

Danach wurden im Überblick „Wie Erklären“ verschiedene Ansätze für Erklärungen in KI-Systemen vorgestellt. Es wurde gezeigt, dass z. B. die

Trainingsdaten zur Erklärung genauso genutzt werden können, wie die konkreten Operationen in der Verarbeitung des Systems. Der letzte Punkt wurde im Kapitel zur „Layerwise-Relevance Propagation" eingehender besprochen und eine konkrete Technologie vorgestellt, die sich mit dem Hervorheben relevanter Informationen – z. B. Pixel – beschäftigt.

Auch auf theoretischer Ebene wurde durch die Vorstellung von Counterfactual Explanations ein Ansatz zur Generierung leicht verständlicher Erklärungen vorgestellt und anhand der Fallbeispiele diskutiert.

Abschließend wurden in diesem Kapitel auch Methoden vorgestellt, die genutzt werden können, um den Einsatz von erklärbaren KI-Systemen in der Praxis zu beurteilen. Dabei wurden subjektive und objektive Messwerte vorgestellt und Einsatzmöglichkeiten dargelegt. Nutzen Sie die Inhalte des Kapitels, um in den zukünftigen Kapiteln zu überlegen, welche Auswirkung das Hinzufügen von Erklärungen z. B. auf die Nachvollziehbarkeit, die erlebte Vertrauenswürdigkeit oder gar die rechtliche Grundlage eines Systems haben (z. B. wenn es um Gleichbehandlung geht).

Literatur

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access, 6,* 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE, 10*(7), e0130140. https://doi.org/10.1371/journal.pone.0130140.

de Winter, J. C. F., Eisma, Y. B., Cabrall, C. D. D., Hancock, P. A., & Stanton, N. A. (2019). Situation awareness based on eye movements in relation to the task environment. *Cognition, Technology & Work, 21*(1), 99–111. https://doi.org/10.1007/s10111-018-0527-6.

Grier, R. A. (2015, September). How high is high? A meta-analysis of NASA-TLX global workload scores. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Bd. 59, No. 1, S. 1727–1731). Sage CA: SAGE Publications.

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for explainable AI: Challenges and prospects. arXiv preprint arXiv:1812.04608.

Jian, J. Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International journal of cognitive ergonomics, 4*(1), 53–71.

Montavon, G., Lapuschkin, S., Binder, A., Samek, W., & Müller, K.-R. (2017). Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognition, 65,* 211–222. https://doi.org/10.1016/j.patcog.2016.11.008.

Montavon, G., Samek, W., & Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing, 73,* 1–15. https://doi.org/10.1016/j.dsp.2017.10.011.

Nushi, B., Kamar, E., & Horvitz, E. (2018, June). Towards accountable ai: Hybrid human-machine analyses for characterizing system failure. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* (Bd. 6, S. 126–135). https://doi.org/10.1609/hcomp.v6i1.13337.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). „Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (S. 1135–1144). https://doi.org/10.1145/2939672.2939778.

Rutjes, H., Willemsen, M., & IJsselsteijn, W. (2019). Considerations on explainable AI and users' mental models. In *Where is the Human? Bridging the Gap Between AI and HCI: Workshop at CHI'19, May 4-9, 2019, Glasgow, Scotland UK.* Association for computing machinery, Inc. http://www.martijnwillemsen.nl/recommenderlab/RutjesChHI2019ws.pdf. Zugegriffen: 15. Okt. 2022.

Schrills, T., Zoubir, M., Bickel, M., Kargl, S., & Franke, T. (2021). Are users in the loop? Development of the subjective information processing awareness scale to assess XAI. In *ACM CHI Workshop Human-Centered Perspectives in Explainable AI*. https://hcxai.jimdosite.com/hcxai-21-papers-and-videos/. Zugegriffen: 15. Okt. 2022.

Sokol, K., & Flach, P. (2018). Conversational explanations of machine learning predictions through class-contrastive counterfactual statements. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence,* 5785–5786. https://doi.org/10.24963/ijcai.2018/836.

Toulmin, S. (2003). Return to reason. In *Return to Reason.* Harvard University Press.

Trommler, D., Attig, C., & Franke, T. (2018). Trust in activity tracker measurement and its link to user acceptance. *Mensch und Computer 2018-Tagungsband.*

Weld, D. S., & Bansal, G. (2019). The challenge of crafting intelligible intelligence. *Communications of the ACM, 62*(6), 70–79.

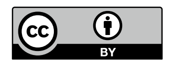
Weiterführende Literatur

Eiband, M., Buschek, D., Kremer, A., & Hussmann, H. (2019). The impact of placebic explanations on trust in intelligent systems. Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, 1–6. https://doi.org/10.1145/3290607.3312787.

Endsley, M. R. (1988). Situation awareness global assessment technique (SAGAT). Proceedings of the IEEE 1988 National Aerospace and Electronics Conference, 789–795. https://doi.org/10.1109/NAECON.1988.195097.

Holzinger, A., Carrington, A., & Müller, H. (2020). Measuring the quality of explanations: The system causability scale (SCS) comparing human and machine explanations. *KI-Künstliche Intelligenz, 34*(2), 193–198. https://doi.org/10.1007/s13218-020-00636-z.

Lapuschkin, S., Binder, A., Montavon, G., Müller, K. R., & Samek, W. (2016). The LRP toolbox for artificial neural networks. The *Journal of Machine Learning Research, 17*(1), 3938–3942.


BY

9 Prozessautomatisierung

Zusammenfassung

Es gibt unterschiedliche Ansätze, Prozesse und Abläufe in öffentlichen Verwaltungen zu verbessern, etwa durch ein kontinuierliches Prozess-Verbesserungsmanagement oder regelmäßige strukturierte Reflexions-Workshops. Aus technischer Sicht können auch hier gegebenenfalls KI-Systeme zum Einsatz kommen. In diesem Kapitel werden entsprechende Einsatzmöglichkeiten vorgestellt. Zunächst wird erörtert, inwiefern Assistenzsysteme Prozesse unterstützen und betroffende Mitarbeitende entlastet werden können (9.2). Anschließend wird ein Überblick über das sogenannte Business Process Management geliefert (9.3). Darauf aufbauend werden Grundlagen zu Robotic Process Automation (9.4) und dessen arbeitspsychologische Auswirkungen bei einem möglichen Einsatz vorgestellt (9.5). Schließlich wird diskutiert, wie sich ein Einsatz von Robotic Process Automation in der Verwaltung gestalten könnte und welche Vor- und Nachteile damit einhergehen (9.6). Das angeeignete Wissen kann in einer Übung reflektiert (9.7) und im eigenen Anwendungsfall eingesetzt werden (9.8).

9.1 Einleitung

Vieles in der öffentlichen Verwaltung ist durch Prozesse geregelt. Vorgänge werden in klar definierten Schritten bearbeitet, mit eindeutigen Zuständigkeiten wer zu welchem Zeitpunkt handeln muss. Dies betrifft nicht nur die eigentliche Sachbearbeitung, sondern auch Kontrollschritte, Zeichnung und Gegenzeichnung, Vier-Augen-Prinzip und Ähnliches. Akten müssen bearbeitet und wieder abgelegt

M. Heine et al., *Künstliche Intelligenz in öffentlichen Verwaltungen*, Edition eGov-Campus, https://doi.org/10.1007/978-3-658-40101-6_9

werden, Vorgangsschritte müssen in einem Protokoll festgehalten werden – zur Qualitätssicherung und späteren Nachvollziehbarkeit.

In dem Maße, in dem Daten und Dokumente solcher Vorgänge digitalisiert werden, bietet es sich auch an, die dazugehörigen Prozesse zu digitalisieren. Doch selbst in einem weitgehend Papier-basierten Vorgang kann eine digitale Prozessbegleitung eine Unterstützung sein. Die Software übernimmt das sogenannte Prozessmanagement, sorgt also dafür, dass Fristen eingehalten werden, benachrichtigt den nächsten Akteur im Prozess und führt automatisch das Verlaufsprotokoll dadurch, wobei alle Beteiligten die Bearbeitung ihres Prozessschrittes quittieren. Diese Information kann wiederum genutzt werden, um Überlastungspunkte und *Hot Spots* über alle Prozesse hinweg zu erkennen und dort einzugreifen.

Der Oberbegriff zu prozessverbessernden Maßnahmen ist Geschäftsprozessmanagement oder auch *Business Process Management* (BPM). Dazu gibt es eine Reihe von Software-Produkten. Um Prozesse zu beschreiben, hat sich die *Business Process Modelling and Notation* (BPMN) als weit verbreiteter Standard etabliert. Diese Prozessbeschreibungen sind die Grundlage für eine Automatisierung, die von einigen Business Process Engines verwendet werden, um die Aktivitäten in einer Prozesskette abzuarbeiten. KI-Systeme können die Abarbeitung solcher Prozesse vielfältig unterstützen. Sie können helfen, Prozessdefinitionen aus existierenden Verlaufsprotokollen herauszuarbeiten, um zunächst feststellen zu können, wie Prozesse überhaupt ablaufen. Sie können aber auch im dynamischen Verlauf eines Prozesses helfen, Fragen zu beantworten wie „Was ist der beste nächste Schritt?“, „Muss jemand benachrichtigt werden?“ oder auch „Wenn es einen Engpass gibt, welcher Vorgang sollte die höhere Priorität bekommen?“.

Ob in der freien Wirtschaft oder in der Verwaltung, die Art und Weise, wie Arbeitsschritte gestaltet werden, ist maßgeblich dafür entscheidend, in welchem Tempo die Arbeit erledigt werden kann. Neben der Effizienz ist auch die Arbeitspsychologie ein wichtiger Faktor. Es gibt schlichtweg unbeliebte Tätigkeiten. Es ist daher ein attraktiver Gedanke, diese Tätigkeiten von einem System erledigen zu lassen. In diesem Kapitel wird es darum gehen, inwieweit RPA die Effizienz von Arbeitsprozessen verbessern und die Zufriedenheit steigern kann. Es werden zunächst Lösungen für Assistenzsysteme vorgestellt. Anschließend wird auf die Grundlagen des BPM und der sogenannten Robotic Process Automation (RPA) eingegangen. Es folgt ein kurzer Einblick in die arbeitspsychologischen Auswirkungen von Prozessautomatisierung. Danach wird der Fokus auf die öffentliche Verwaltung gesetzt und die Relevanz der angesprochenen Technologien wie etwa RPA diskutiert.

9.2 Assistenzsysteme

Ein Großteil der Prozesse in der öffentlichen Verwaltung beginnt mit einer ersten Kontaktaufnahme. Das kann über ein Formular geschehen, welches digital oder direkt vor Ort abgegeben wird, ein Telefonat oder das persönliche Erscheinen in einer Behörde, um das jeweilige Anliegen vorzubringen. Assistenzsysteme, die eine solche Kontaktaufnahme unterstützen, sind beispielsweise Chatbots auf Behörden-Webseiten, Service-Roboter vor Ort oder Sprachassistenten, die beim Ausfüllen von Formularen unterstützen. Dadurch werden der Zugang zu relevanten Informationen erleichtert, die Fehlerhäufigkeit und die Bearbeitungszeit reduziert und somit insgesamt Prozesse verbessert. In der Regel ist eine der ersten Anforderungen an solche Systeme, dass sie in der Lage sind mit natürlicher Sprache zu interagieren.

Beispiel

Es gibt bereits einige Assistenzsysteme, die in deutschen Behörden im Einsatz sind. Der Service-Roboter L2B2 aus Ludwigsburg ist ein Beispiel für ein Assistenzsystem vor Ort, welches quasi die Rezeption oder eine Pforte ersetzt. Technologisch besteht große Ähnlichkeit zu einem Chatbot auf einer Webseite, nur dass hier der Chatbot eine physische Gestalt in Form eines Roboters erhält. Personen, welche das Gebäude betreten, werden begrüßt, bekommen Informationen zu den im Rathaus angebotenen Leistungen, erfahren, welche Unterlagen benötigt werden und können nach dem Weg zur richtigen Abteilung fragen. Die Robotergestalt ermöglicht zusätzliche Barrierefreiheit, da L2B2 beispielsweise Menschen mit Seheinschränkungen direkt zur Tür der zuständigen Abteilung führen kann (BMWi, 2020). Die Stadt Ludwigsburg setzt neben L2B2 noch ein weiteres Assistenzsystem zur Digitalisierung der Behördengänge ein: eine Service-Station, an der für den Reisepass oder den Personalausweis biometrische Passbilder erstellt oder Fingerabdrücke erfasst werden können. Auch die Anträge können schon digital vorausgefüllt werden. Hierdurch wird die Wartezeit für die antragstellende Person verkürzt und Mitarbeitenden der Behörde werden von Routinetätigkeiten entlastet.◄

Die Digitalisierung der Verwaltung wird im Zuge des Onlinezugangsgesetzes (OZG) deutlich vorangetrieben und dieser Anwendungsfall aus der Stadt Hamburg ist ein sehr gutes Beispiel dafür, wie mehrere Prozesse, die eng miteinander

verknüpft sind, neu gedacht und zum Vorteil aller Beteiligten neu gestaltet werden können.

Beispiel

Für die Lebenslage „Geburt eines Kindes" wurde gemeinsam mit Geburtskliniken, Standesämtern und der Kindergeldstelle ein institutionsübergreifender Geschäftsprozess entwickelt. In einem ersten Schritt wurden mehrere Einzelprozesse wie die Anmeldung des Kindes beim Standesamt und die Beantragung von Kindergeld in einem kombinierten Formular zusammengeführt, sodass Eltern die nötigen Daten und die entsprechenden Nachweise nur noch einmal angeben und vorlegen müssen, die beteiligten Institutionen kümmern sich um alles Weitere. Eltern erhalten so deutlich schneller die Geburtsurkunde des Kindes, die Steuer-ID, den Eintrag im Melderegister und den Kindergeld-Bescheid per Post und sparen Wege und bürokratischen Aufwand, insbesondere in der so besonderen Zeit wenige Tage nach der Geburt ihres Kindes.

Das Projekt ist mittlerweile so weit fortgeschritten, dass in allen Geburtskliniken in Hamburg der neue Prozess auch digital etabliert ist und die Antragsbearbeitung der Eltern durch einen Sprachassistenten unterstützt wird. Der Sprachassistent, der beim Ausfüllen des neuen Kombi-Formulars hilft, steht entweder über Service-Terminals in den Kliniken oder über eine Web-Version mittels Smartphone oder Tablet zur Verfügung. Der Einsatz des Sprachassistenten und die digitale Antragstellung erhöhen die Barrierefreiheit für Menschen mit Sehbehinderungen und Mobilitätseinschränkungen. In einer Erweiterung des Assistenten wäre auch eine automatische Übersetzungsfunktion denkbar, die dann insbesondere Menschen mit anderen Muttersprachen und geringen Deutschkenntnissen den Zugang zu den Verwaltungsleistungen erleichtern (Stadt Hamburg, 2022).◄

9.3 Business Process Management

Die in einer Behörde gültigen Prozesse haben einen wesentlichen Einfluss auf die auszuführenden Aufgaben, die zugehörigen Verantwortlichkeiten und nicht zuletzt auf die alltägliche Arbeit der Beschäftigten. Prozesse zeigen die Interaktionen und Informationsflüsse zwischen unterschiedlichen Abteilungen und

Beteiligten auf, häufig werden dazu verwendete IT-Systeme und während des Prozessablaufes erstellte Artefakte aufgezeigt. Verschiedene Prozesse tragen zu einer Verwaltungsleistung bei. Durch die Beschreibung von Prozessen wird deutlich, welche Aktivitäten von welchen Abteilungen erledigt werden. Aktivitäten können dabei automatisierte Aufgaben sein oder manuelle, durch Beschäftigte getätigte Schritte. Diese Informationen bilden die Grundlage festzustellen, zwischen welchen Übergängen von Aktivitäten Medienbrüche (Papier/digital) offensichtlich werden – dabei ist eine medienbruchfreie Integration das Ziel. Analoge Prozesse sollten nicht 1-zu-1 digital abgebildet und automatisiert werden, vielmehr sollte bei solchen Transfers die Chance genutzt werden Prozesse zu optimieren. Die Tätigkeiten rund um das Planen, die korrekte Ausführung, die Optimierung oder das Steuern und Messen von Prozessen werden unter dem Begriff *Business Process Management* (BPM) zusammengefasst (vgl. dazu Becker et al., 2009). Das BPM konzentriert sich nicht nur auf Verbesserungen einzelner Bereiche, sondern strebt eine ganzheitliche Betrachtung von Prozessen in einer Organisation an. Im behördlichen Kontext wird auch häufig von einem Vorgang statt von einem Prozess gesprochen.

Es gibt bereits eine Vielzahl an Referenzmodellen und Frameworks für das BPM, etwa die *Information Technology Infrastructure Library* (ITIL) für IT-relevante Services (vgl. Schaaf, 2007), *Six Sigma* für das Qualitätsmanagement (vgl. Arcidiacono et al., 2012) oder das *Supply Chain Operation Reference Modell* (SCOR) zum Informationsaustausch zwischen Unternehmen in einer Lieferkette (vgl. Ahoa et al., 2018). In diesen Modellen sind bereits zahlreiche Prozesse und Indikatoren beschrieben – teilweise so detailliert, dass diese unmittelbar angewendet werden können.

Das zentrale Ziel des BPM ist es sicherzustellen, dass Prozesse effektiv und effizient durchgeführt werden. Des Weiteren soll sichergestellt werden, dass der Ressourcenverbrauch, die Durchlaufzeiten und die Anzahl der Organisationsschnittstellen möglichst reduziert werden. Außerdem soll sichergestellt werden, dass Prozesse angemessen überwacht und gesteuert sowie kontinuierlich optimiert werden. Schließlich soll auch die Qualität des Prozessergebnisses fortlaufend überprüft werden (vgl. Huber & Huber, 2011, S. 4).

Man unterscheidet zwischen sogenannten *Ist*-Prozessen – Prozesse, wie sie aktuell durchgeführt werden – und *Soll*-Prozessen, also der Beschreibung zukünftiger Prozesse und zugehöriger Handlungsmaßnahmen, um diese zu erreichen. Prozesse werden in definierten Modellierungssprachen abgebildet. Eine dieser semi-formalen Sprachen, die in der Praxis häufig zum Einsatz kommt und mittlerweile als Standard gilt, ist die *Business Process Model and Notation* (BPMN).

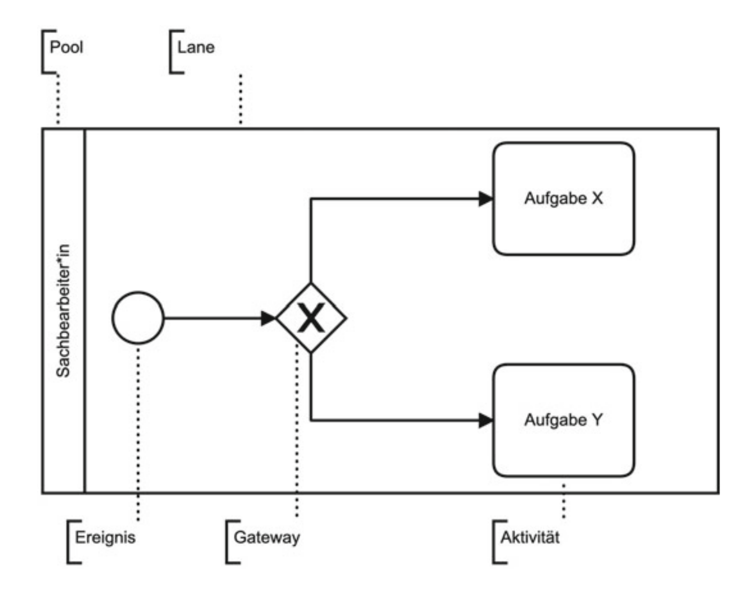


Abb. 9.1 Ausgewählte Elemente des BPMN

Häufig genutzte Elemente sind Flussobjekte, wie beispielsweise eine Aktivität, ein Ereignis oder ein Gateway sowie verbindende Objekte, etwa ein Pfeil, durch den der Ablauf nachvollziehbar wird (siehe Abb. 9.1). Die waagerechten Linien bilden sogenannte Pools und Lanes (Bahnen), wodurch die verschiedene Teilnehmende im Prozess dargestellt werden.

Spezielle Programme, sogenannte Business Process Engines, sind in der Lage, eine Prozessbeschreibung abzuarbeiten. Sie starten an einem ausgewiesenen Startschritt, führen die automatisierten Aufgaben aus oder warten darauf, dass manuelle Aufgaben getätigt werden, und gehen dann zum nächsten Prozessschritt. Der Vorgang wird wiederholt, bis der Prozess endet. Auch hierfür gibt es bei BPMN eine spezielle Notation.

Ob sich der jeweilige Prozessschritt automatisiert erledigen lässt, muss individuell geprüft werden. Eine vorgefertigte E-Mail zu versenden, z. B. mit einer Benachrichtigung, dass bald eine Frist verstreicht, ist verhältnismäßig einfach umzusetzen. Eingehende elektronische Dokumente zu erfassen und an die richtige Abteilung weiterzuleiten, kann schon herausfordernder sein, da die Prüfung von Anträgen in der Regel den Beschäftigten vorbehalten ist. Trotzdem kann eine Prozessautomation hier dafür sorgen, dass die Beendigung der Prüfung registriert

und ein nachfolgender Prozessschritt angestoßen wird, z. B. die Erstellung eines Bescheides, indem die Process Engine die dafür zuständige Stelle benachrichtigt.

Oft sind gerade die Entscheidungspunkte im Prozessplan die Stellen, bei denen ein Mensch in der Regel eingreifen muss. Zum Beispiel wird eine Mitarbeiterin informiert, dass ein Prozess auf eine Entscheidung wartet, die in ihrem Zuständigkeitsbereich liegt. Am Rechner könnte sich ein Dialog öffnen, der die notwendige Entscheidung abfragt, erst danach kann der Prozess fortgesetzt werden. Hängen diese Entscheidungen von klaren Regeln ab, die sich auf objektive, überprüfbare Fakten stützen, zum Beispiel ob ein bestimmtes Datum noch mindestens 10 Tage in der Zukunft liegt, ob eine Antragssumme unter einem vorgegebenen Betrag liegt oder ob die Liste der Belege am aktuellen Vorgang vollständig ist, liegt die Überlegung nahe, diese Überprüfungen automatisch durchzuführen. In solchen Fällen, wo die Fakten elektronisch überprüfbar sind, ist es sinnvoll, die entsprechende Entscheidung zu automatisieren. Dazu ist es nötig, die Regeln zu hinterlegen, sodass die Prozessverarbeitung direkt überprüft und der dazugehörige nächste Prozessschritt initiiert werden kann.

9.4 Grundlagen zu Robotic Process Automation

Im Gegensatz zu anderen Technologien gilt die Einführung von Prozessautomatisierung mittels *Robotic Process Automation* (RPA) als schnell und vergleichsweise einfach, gleichzeitig wird der Kostenaufwand als verhältnismäßig gering eingeschätzt. In der Privatwirtschaft kommen RPA-Systeme bereits häufig zum Einsatz, insbesondere mit dem Ziel, Prozesskosten zu minimieren. Es gibt mittlerweile eine Vielzahl von Anbietern für RPA-Systeme und auch Beratungshäuser haben bereits begonnen, sich auf RPA zu spezialisieren. Im Jahr 2019 wurde der Markt von etwa 50 Anbietern für RPA abgedeckt (Smeets et al., 2019, S. 49). RPA-Software wird schätzungsweise bereits bei circa 30 % der im Deutschen Aktienindex gelisteten Unternehmen eingesetzt und in den kommenden Jahren wird eine deutliche Zunahme erwartet (Reich & Braasch, 2019, S. 302). RPA hat den größten Durchbruch in der Qualitätssicherung in der Softwareentwicklung. Fehlerhafte Softwareentwicklungen werden der Programmiererin oder dem Programmierer in dem Moment der Fertigstellung durch das RPA-System zurückgemeldet.

Nach Botar et al. (2018, S. 1) handelt es sich bei RPA um eine „robotergesteuerte Prozessautomatisierung“. Diese Roboter treten in Form von Programmsoftware auf, die auf Computern installiert und selbstständig in bestehenden Anwendungen tätig ist. Dabei interagieren die Roboter über die Anwender-Ebene

mit anderen Systemen und imitieren dabei menschliches Handeln. Diese Roboter sind also keine Maschinen im physischen Sinne, sondern Programme, die Mitarbeitende bei der Erledigung ihrer Aufgaben assistieren oder sie bei bestimmten Tätigkeiten ersetzen (Allweyer, 2016, S. 1). Nach van der Aalst (2018, S. 269) dient RPA als ein Oberbegriff für Software-Tools, die wie ein Mensch über die Anwender-Oberfläche des Computers auf andere Anwendungen zugreifen und dabei Änderungen durchführen. Diese Programme zielen darauf ab, den Menschen bei der Erledigung ihrer Aufgaben zu helfen oder einzelne Tätigkeiten zu übernehmen. Somit ist RPA in der Lage, Arbeitsprozesse teilweise oder vollständig eigenständig durchzuführen. Es handelt sich um einen Lösungsansatz, für den im Grunde keine wesentlichen Anpassungen in der bestehenden IT-Architektur vorgenommen werden müssen. Das Programm bedient die Benutzerschnittstelle genauso, wie es ein Mitarbeiter tun würde. Der Roboter meldet sich im System mit seinen Benutzerdaten an und führt die jeweiligen Tätigkeiten in den Anwendungen aus. RPA lässt sich deshalb als non-invasive Technologie definieren. Aufgrund der Emulation der Eingaben auf der vorhandenen Anwendungsoberfläche müssen bestehende Anwendungen nicht geändert werden (Czarnecki & Auth, 2018, S. 116). Häufig wird auch von Bots statt von Robotern gesprochen, um deutlich zu machen, dass es sich um eine automatisiert handelnde Softwarelösung handelt (Smeets et al., 2019, S. 7 ff.). RPA ist für alle Prozesse geeignet, die eindeutig geregelt und an klaren Standards ausgerichtet sind. Dazu zählen etwa die Bearbeitung von großvolumigen Prozessvorgängen und Arbeitsabläufen. Die Arbeit eines RPA-Bots ist in hohem Maße von Effizienz geprägt. Tätigkeiten, die sich automatisiert ausführen lassen, erledigt der Roboter zu sehr geringen Kosten und in einer deutlich kürzeren Zeit als der Mensch. Neben der Verringerung der Kosten wird eine höhere Qualität erwartet. So entfallen menschliche Fehler durch Ablenkung, Müdigkeit und Krankheit, und der Bot kann pausenlos 24 h pro Tag eingesetzt werden. Prozesse, die automatisiert ablaufen, können vom Anfang bis zum Ende überprüft werden und erreichen einen hohen Präzisionsgrad. RPA übernimmt die monotonen Routine-Tätigkeiten, wodurch sich das Personal fortan in dieser Zeit auf Bereiche konzentrieren kann, in denen eine höhere Konzentration gefordert ist (Reich & Braasch, 2019, S. 297). Darüber hinaus ist RPA-Software im Unterschied zu traditionellen Lösungen in der Lage zu lernen und standardisierte Vorgänge eigenständig zu bearbeiten. In diesen Fällen spricht man von *Intelligent Process Automatisation* (IPA). RPA ist nun mit einem KI-System angereichert, welches auf Grundlage der bisherigen Prozessdaten fortlaufend lernt (Reich & Braasch, 2019, S. 296).

Bei der Auswahl passender Prozesse spielen einerseits wirtschaftliche Aspekte eine Rolle, andererseits sind technische Faktoren entscheidend. Zu Ersterem

gehören etwa die Einsparung von Kosten, die Erhöhung der Prozessqualität oder die Verringerung des Zeitaufwands für einzelnde Mitarbeitende. Welches Prozessvolumen für den wirtschaftlichen Einsatz von RPA als sinnvoll erachtet wird, ist umstritten. Ein hohes Prozessvolumen führt in der Regel zu großen Einsparungen. Durch den geringen Implementierungsaufwand kann allerdings ein Einsatz von RPA auch bei kleineren Prozessen bereits gerechtfertigt sein.

Zu den technischen Prozess-Auswahlkriterien gehören etwa der Grad der Standardisierung (denn je höher dieser ist, desto besser lässt sich ein Prozess automatisieren), die Regelbasiertheit (nur regelbasierte Prozesse können vollständig automatisiert werden) oder die Stabilität des Prozesses (wenn sich ein Prozess häufig ändert, müssen diese Änderungen auch im RPA-Ablauf angepasst werden, was ggf. zu einem unverhältnismäßig hohem Aufwand führt). Selbstverständlich müssen die Prozessdaten in digitaler Form vorliegen und es müssen ferner strukturierte Daten sein. Außerdem gilt, je höher die Anzahl der beteiligten Anwendungen, die im Prozess durchlaufen werden, desto eher lohnt sich der Einsatz von RPA (vgl. für die genannten Auswahlkriterien auch Smeets, 2019, S. 40).

Chancen und Herausforderungen von RPA

Chancen von RPA

- *Kostenreduktion:* Eines der am häufigsten genannten Argumente für die Automatisierung von Prozessen ist die Reduzierung der Kosten. Es wird teilweise geschätzt, dass für den Einsatz von RPA nur ein Neuntel der Kosten anfallen im Vergleich zur manuellen Bearbeitung (vgl. Deloitte, 2015, S. 7).
- *Qualitätssteigerung:* Insbesondere bei standardisierten und repetitiven Aufgaben neigen Menschen zu Fehlern, weil die Konzentration sinkt. Ein Bot arbeitet hingegen nahezu fehlerfrei, was wiederum zu einer höheren Qualität des Prozesses führt (Smeets et al., 2019, S. 24). RPA ermöglicht es zeitgleich, die Prozesstransparenz zu erhöhen, weil beispielsweise automatisierte Reports erstellt werden. Hierdurch können frühzeitig mögliche Prozessoptimierungen eingeleitet werden (vgl. Allweyer, 2016, S. 5).
- *Zeiteinsparung:* In Korrelation zur Einsparung von Kosten steht der Faktor Zeit. Eine geringere Bearbeitungszeit bindet weniger Kapazitäten und Produktionsmittel, wodurch sich dann auch die Kosten reduzieren.

Doch nicht nur die damit einhergehende Kostenreduktion ist als Vorteil der Zeiteinsparung zu sehen, auch die schnellere Prozessabwicklung kann vorteilhaft sein, insbesondere bei Prozessen, in denen Bürgerinnen und Bürger miteinbezogen sind. Denn eine höhere Prozessabwicklung steigert die Zufriedenheit, etwa im Antragswesen.

Herausforderungen von RPA

- *Instandhaltung:* Für die Implementierung von RPA müssen zunächst Ressourcen bereitgestellt werden. Aber auch nach der Implementierung muss das System aktualisiert und regelmäßig überprüft werden. Auch hierfür müssen entsprechende Ressourcen eingeplant werden.
- *Geringe Stabilität:* RPA-Lösungen sind im Vergleich zur Automatisierung über Schnittstellen weniger stabil. Über Schnittstellen lässt sich eine umfassende Systemintegration erreichen. Dies ist bei RPA-Systemen nicht der Fall, da sie über die graphische Benutzeroberfläche arbeiten. Diese Oberfläche kann sich aber regelmäßig ändern, dies hat dann Auswirkungen auf den mit RPA unterstützten Prozess. Bei Schnittstellen hat dies keine Bedeutung, da sie im Backend arbeiten.

9.5 Arbeitspsychologie und RPA-Einsatz

Was beeinflusst den Erfolg einer Tätigkeit? Wenn ein Mensch sich mit einer Tätigkeit beschäftigt, so spielen maßgeblich folgende Faktoren eine Rolle:

- die Geschwindigkeit, mit der die Tätigkeit ausgeführt wird;
- das möglichst fehlerfreie Ausführen der Tätigkeit;
- die Aufmerksamkeit bzw. der notwendige Konzentrationsgrad (Wickens, 2016).

Es soll demnach möglichst schnell und möglichst fehlerfrei gearbeitet werden. Beide Aspekte werden vom dritten Punkt beeinflusst. Erfordert eine Tätigkeit hohe Konzentration, damit keine Fehler entstehen, so wird der ausführende Mensch gleichzeitig kaum eine hohe Geschwindigkeit aufrechterhalten können. Erwachsene können sich durchschnittlich maximal 25 min am Stück konzentrieren, bevor eine Pause benötigt wird. Die Exaktheit leidet ebenfalls unter

einer hohen Aufmerksamkeitsforderung. Sinkt das Konzentrationsniveau zwischenzeitlich auch nur minimal ab, so kann schon ein Fehler passieren. Einfache Tätigkeiten, die„nebenbei“ erledigt werden können, sind weniger fehleranfällig als anspruchsvolle. Allerdings gibt es auch hier eine wichtige Ausnahme. Eigentlich banale Aufgaben, die sich aber ständig wiederholen, werden auf Dauer zur Belastung. Repetitive Tätigkeiten schmälern die neuronale Aktivität im Gehirn. Es kommt auf Dauer zu Ermüdungszuständen, dementsprechenden Konzentrationsstörungen und letztendlich leidet die Exaktheit darunter. Bei der Anwendung von RPA kann vor allem die zuletzt genannte Arbeitsbelastung reduziert werden. Formalitäten, die häufig vorkommen und einem Muster entsprechen, können von RPA abgehandelt werden. Rückblickend auf die drei Faktoren kann auf diese Weise die Akkuratheit bei sich wiederholenden Prozessen gesteigert werden. RPA ist der erste Schritt zur Automatisierung, die kostengünstig und zügig implementiert werden kann. Die Alternative, nämlich durch eine feste Programmier-Schnittstelle (API) im Backend Daten zu ändern, ist langfristig im Wartungsaufwand geringer. Abhängigkeiten, die durch die Manipulation über die graphische Oberfläche mit RPA entstehen, werden durch APIs vermieden – die Software-Architektur bleibt weniger komplex. Es muss daher vorab eine genaue Priorisierung bestimmt werden, welches Ziel mithilfe von Automatisierung verfolgt werden soll.

9.6 Zum Einsatz von RPA in der Verwaltung

Viele Prozesse in der öffentlichen Verwaltung erfordern Datenübertragungen von A nach B. So kann es beispielsweise erforderlich sein, Daten von einer Papierrechnung in eine Finanzbuchhaltungssoftware zu übertragen oder Stammdaten aus dem einen System in das andere zu übertragen. Teilweise sind die entsprechenden Fachverfahren veraltet und es mangelt an Möglichkeiten des Datenaustauschs, sodass die Datenübertragungen händisch vom Personal erfolgen müssen. Damit verbunden ist oftmals eine starke Monotonie und es kann zu Ermüdungszuständen und Konzentrationsstörungen der Beschäftigten kommen – insbesondere wenn die Aufgabe über einen längeren Zeitraum ausgeführt wird. In der Folge kann es zu fehlerhaften Datenübertragungen kommen. Doch gerade bei Aufgaben wie diesen handelt es sich meist um Prozesse, die stark regelbasiert und standardisiert sind. Es liegt nahe, an dieser Stelle digitale Unterstützung – wie etwa RPA – heranzuziehen, um Beschäftigte zu entlasten und Ressourcen neu verteilen zu können.

Grundsätzlich lassen sich Bots ohne Programmierkenntnisse erstellen. Viele RPA-Anbieter bieten grafische Benutzeroberflächen zum Erstellen von Bots an, in denen die Bots per Drag-and-Drop „programmiert“ werden können. Auch mit sogenannten Makro-Recordern lassen sich Bots erstellen, indem die Interaktionen mit einer Anwendung auf dem Bildschirm aufgezeichnet werden. Bekannte Firmen in diesem Segment sind UiPath, Automation Anywhere, Blue Prism oder WorkFusion.

Allgemein gilt: Je fester das Regelwerk bzw. je höher der Standardisierungsgrad eines Prozesses ist, desto höher ist auch die Sicherheit, dass ein Bot fehlerfrei funktioniert. Steigt die Anzahl variabler Größen im Prozess, steigt auch die Fehleranfälligkeit bzw. die Wartungskomplexität des Bots. Die Vorteile eines RPA-Einsatzes können wie folgt zusammengefasst werden:

- Entlastung von Beschäftigten und freie Ressourcen für andere Aufgaben verfügbar;
- schnelle Aufgabenausführung;
- hohe Verarbeitungsqualität bei festem Regelwerk;
- Bots arbeiten rund um die Uhr;
- meist geringer Aufwand für die Einrichtung.

Aber ist RPA aufgrund dieser Vorteile nun uneingeschränkt zu empfehlen? Auch wenn RPA viele wertvolle Vorteile bietet, lässt sich diese Frage nicht pauschal beantworten. Es sollte immer genau geprüft werden, ob Prozessautomatisierung für das jeweilige Szenario einen Mehrwert bietet. Denn neben den oben genannten Vorteilen bringt RPA auch einige Nachteile bzw. Herausforderungen mit sich, die bei der Entscheidung für den Einsatz dieser Technologie berücksichtigt werden sollten:

- zusätzliche Systemabhängigkeiten, Komplexität, Wartungsaufwände;
- Verantwortlichkeitsübertragung von Mensch zu Maschine (Wie viel kann einem Bot „zugemutet“ werden?);
- Aufgaben, die Bewertungen oder Kreativität erfordern, können (noch) nicht oder nur teilweise automatisiert werden.

Fallbeispiel 1: Der Prozess zur Ausstellung eines Anwohnerparkausweises.

Bei vielen Verwaltungsentscheidungen kann das Ergebnis unmittelbar aus den relevanten Gesetzestexten abgeleitet werden, ohne dass eine weitere Interpretation des Sachverhalts notwendig oder ein Ermessensspielraum vorhanden ist. Der Prozess beinhaltet lediglich die Überprüfung gesetzlich vorgegebener Voraussetzungen. So muss beispielsweise bei der Ausstellung eines Anwohnerparkausweis geprüft werden, ob die antragstellende Person die Voraussetzung für einen solchen Ausweis erfüllt. In diesem Fall beinhaltet dies, dass die Person einen Wohnsitz oder ein Gewerbe in dem jeweiligen Gebiet haben muss. Ein manueller Prozess könnte also wie in Abb. 9.2 aussehen.

Die Aktivitäten mit einem Plus-Zeichen verweisen auf einen Unterprozess, hinter der Aktivität „Zuweisungsprozess der Behörde" stecken also weitere Aktivitäten, auf dessen Darstellung hier zur besseren Übersicht verzichtet wird. Wichtig ist nur zu wissen, dass es ein Verfahren gibt, durch das der spezifische Antrag an eine bestimmte Stelle in der Verwaltung zur Bearbeitung übergeben wird.

Durch eine Schnittstelle zum Einwohnermelderegister kann die manuelle Entscheidungsaktivität über Ablehnung oder Ausstellung des Parkausweises automatisiert werden. Aber wie kann man diese Automatisierung erreichen? Man kann beispielsweise RPA einsetzen. Der digitale Roboter lernt über Beobachtung der Anwenderoberfläche die Prozessschritte, die von ihm in Zukunft ausgeführt werden sollen. Der Prozess könnte dadurch wie in Abb. 9.3 aussehen (Voraussetzung ist hierbei, dass der Antrag digital gestellt wird.):

RPA kommt insbesondere bei einfachen Prozessen zum Einsatz. Man kann diese Art der Automatisierung jedoch mit KI verbinden, um auch bei komplizierteren Prozessen Aufgaben automatisiert abzuarbeiten. Bezogen auf den Use Case kann eine durch KI erweiterte RPA auch mit Fällen umgehen, in denen beispielsweise der Antrag fehlerhaft oder unvollständig ist.◄

Fallbeispiel 2: Einsatz von Robotic Process Automation als Digitalisierungsbrücke

Für die Beantragung von Beihilfen wurde ein digitales Web-Portal geschaffen, als Ablösung des analogen Papierprozesses. Das eigentliche Fachverfahren zur Abwicklung des Beihilfeprozesses konnte jedoch aufgrund verfügbarer Ressourcen und technischer Komplexität nicht erneuert oder erweitert werden. Die Anwendung verfügt über keinerlei Schnittstellen, sodass eine Systemkopplung nicht möglich ist. Die rund 2500 wöchentlich eingehenden

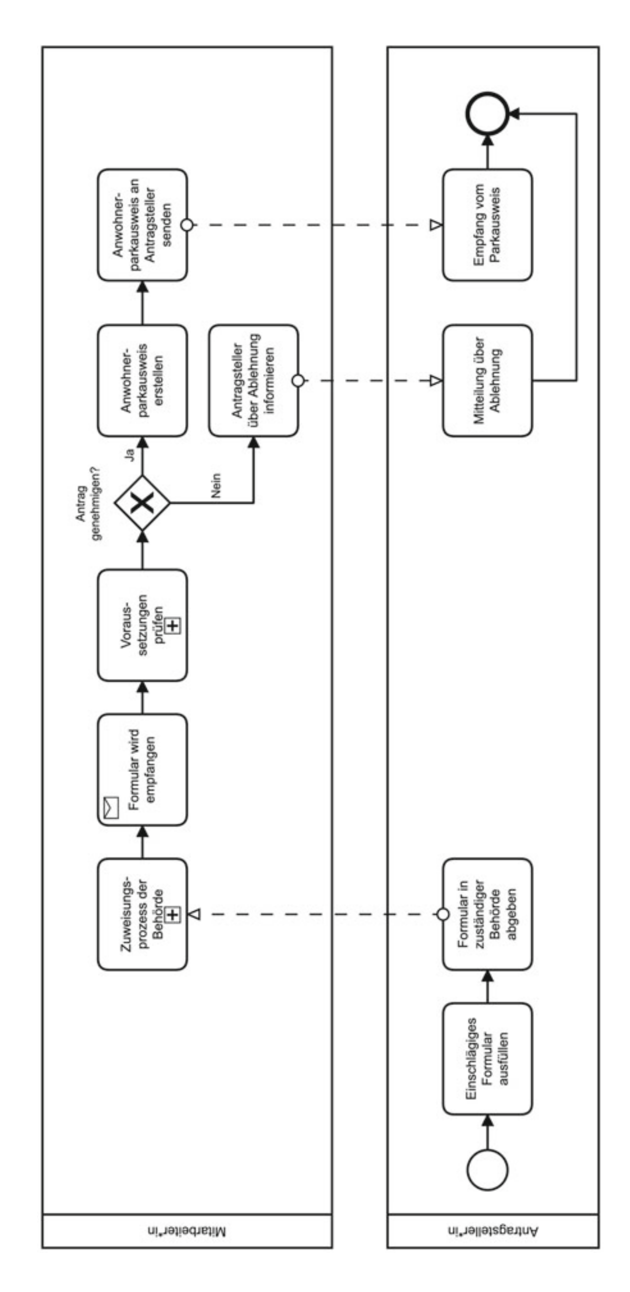


Abb. 9.2 Prozess zur Ausstellung eines Anwohnerparkausweises

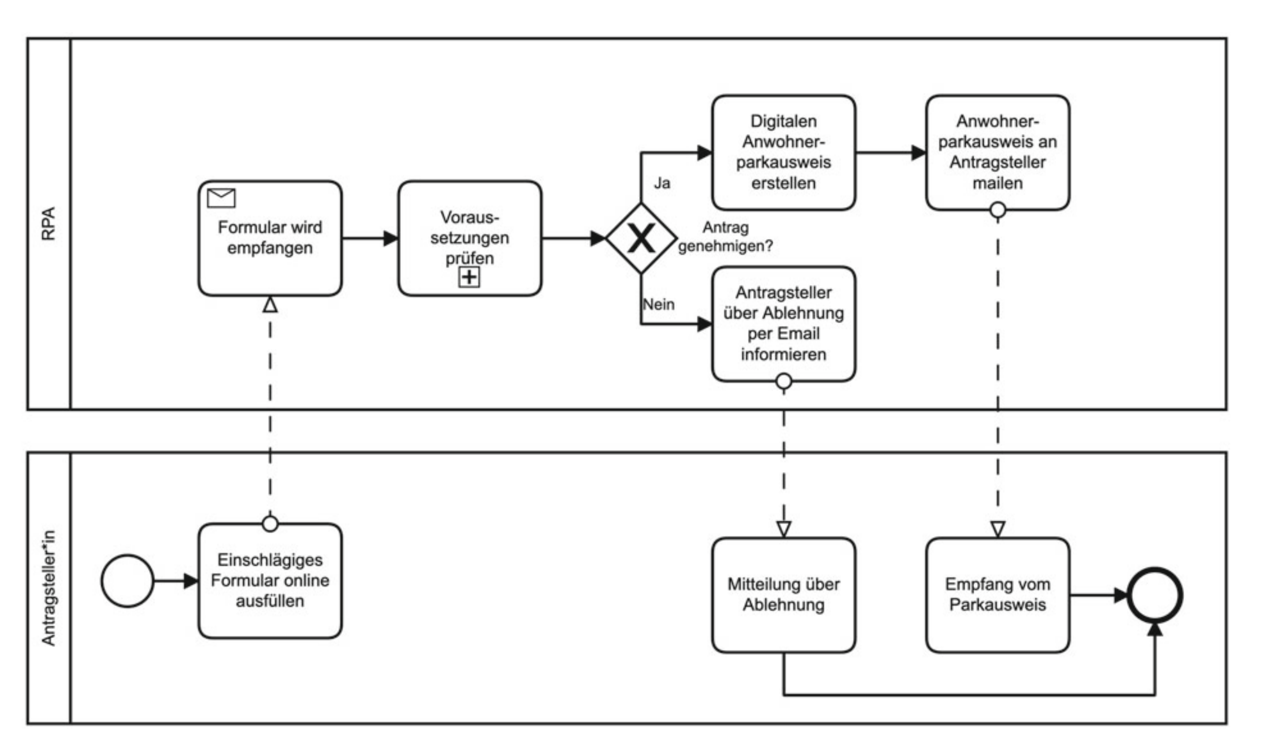


Abb. 9.3 RPA-unterstützter Prozess zur Ausstellung eines Anwohnerparkausweises

Beihilfeanträge mussten somit durch Mitarbeitende von einem in das andere System übertragen werden. Eine Analyse des Prozesses ergab, dass pro Durchlauf ca. 3 min an Arbeitszeit anfallen. Diese Bearbeitungszeit berücksichtigt nicht die Fehlerrate sowie anfallende Aufwände für Nacharbeiten. Bei rund 125 Arbeitsstunden pro Woche ergibt sich daraus ein Arbeitsbedarf von rund 4 Personen. Zudem handelt es sich hierbei um eine sehr monotone und unliebsame Tätigkeit. Durch den Einsatz von Robotic Process Automation konnte eine Lösung geschaffen werden, die eine weitestgehende Automatisierung des Prozesses ermöglicht. Der Roboter prüft das Web-Portal auf eingehende Beihilfeanträge, extrahiert die Daten und überträgt sie in das Fachverfahren. Die Durchlaufzeit pro Vorgang konnte auf 30 s reduziert und die Fehlerrate auf unter 1 % gesenkt werden. Die Komplexität bei der Automatisierung des Prozesses lag vor allem in der Fehlerbehandlung (z. B. Umgang mit inkonsistenten Daten). Anträge, die der Roboter nicht automatisiert bearbeiten kann, werden an einen Mitarbeiter weitergeleitet, der die Anträge gesondert prüft.◄

9.7 Übung zur Prozessoptimierung

1. Die Abkürzung RPA steht für?
   a) Risiko- und Problem-Analyse
   b) Rapid Prototyping Act
   c) Robotic Process Automation
   d) Richtig Prozesse abarbeiten
2. Welche Aussagen zum Service „Kinderleicht zum Kindergeld" der Hansestadt Hamburg sind zutreffend?
   a) Einzelprozesse – wie die Anmeldung des Kindes beim Standesamt oder die Beantragung von Kindergeld – wurden in einem kombinierten Formular zusammengefasst, sodass Eltern die notwendigen Daten nur noch einmal vorlegen müssen.
   b) In Hamburger Geburtskliniken kann das Formular an einem Service-Terminal ausgefüllt werden. Dort wird man außerdem durch einen Sprachassistenten unterstützt.
   c) Das Ausfüllen des Formulars kann von Siri erledigt werden.
   d) Die Eltern müssen lediglich per E-Mail ihre Kontodaten an die zuständige Behörde senden und schon erhalten sie für ihr Kind Kindergeld.

3. Welche Aussagen über das Business Process Management sind zutreffend?
   a) BPM zielt alleinig darauf ab, den Menschen durch Programme und IT-Systeme zu ersetzen und dadurch die Kosten zu reduzieren.
   b) Es gibt bisher kaum Referenzmodelle und Frameworks, auf die man zur Implementierung von BPM zurückgreifen könnte.
   c) BPM zielt darauf ab, die Prozessqualität zu erhöhen sowie die Anzahl der Fehler und die Durchlaufzeit zu reduzieren.
   d) Das BPM konzentriert sich nicht nur auf die Verbesserungen einzelner Bereiche, sondern strebt eine ganzheitliche Betrachtung von Prozessen innerhalb einer Organisation an.
   e) Es gibt bereits zahlreiche Referenzmodelle und Frameworks, die man zur Implementierung von BPM nutzen kann.

   Welche der folgenden Aspekte müssen bei der Auswahl von Prozessen, die für RPA geeignet sind, berücksichtigt werden?
   a) Soziale Aspekte wie etwa, ob der Prozess bei der Belegschaft beliebt ist oder nicht.
   b) Politische Aspekte wie etwa Signale von politischen Entscheidungsträgern.
   c) Wirtschaftliche Aspekte wie etwa die Reduzierung von Kosten.
   d) Technische Aspekte wie etwa die Strukturiertheit der Daten.

4. Mit welcher Sprache werden Geschäftsprozesse modelliert?
   a) Business Process Language of Execution (BPLE)
   b) Business Process Notation and Modelling (BPNM)
   c) Business Process Modelling and Notation (BPMN)
   d) Language of Business Process Notation (LBPN)
5. Wofür wird RPA verwendet?
   a) RPA dient der Automatisierung von Eingaben auf grafischen Oberflächen eines Computersystems.
   b) RPA dient der automatisierten Ausführung aller Aktivitäten eines Prozesses.
   c) RPA hilft der Anwenderin oder dem Anwender eines Software-Systems schneller die Flächen zu finden, die angeklickt werden sollen.
   d) RPA sind Schnittstellen im Backend von Computersystemen (bspw. werden damit Synchronisationen von Datenbanken ermöglicht).

9.8 Aufgaben zum eigenen Anwendungsfall

- Ihr KI-System wird bestehende Prozesse beeinflussen. Skizzieren Sie daher kurz mit einer Modellierungssprache Ihrer Wahl (z. B. BPMN), wie der Prozess derzeit aussieht (Ist-Zustand) und wie sich der Prozess durch Ihr KI-System verändert (Soll-Zustand).
- Überlegen und begründen Sie außerdem, ob der unterstützende Einsatz von RPA für Ihren Anwendungsfall sinnvoll ist.

Literatur

Ahoa, E., Kassahun, A., & Tekinerdogan, B. (2018). Configuring supply chain business processes using the SCOR reference model. *BMSD, 2018,* 338–351. https://doi.org/10.1007/978-3-319-94214-8_25.

Allweyer, T. (2016). Robotic process automation–Neue Perspektiven für die Prozessautomatisierung. *Kaiserslautern: Hochschule Kaiserslautern, 3,* 2019.

Arcidiacono, G., Calabrese, C., & Yang, K. (2012). Leading processes to lead companies: Lean Six Sigma. *Springer Milano.* https://doi.org/10.1007/978-88-470-2492-2.

Becker, J., Mathas, C., & Winkelmann, A. (2009). *Geschäftsprozessmanagement*. Springer.

Botar, A., Pletschacher, M., & Stummeyer, C. (2018). Die Roboter sind da – Wie Robotic Process Automation (RPA) Arbeitnehmer entlastet und Arbeitgebern hohe Kosten spart. *Controller Magazin, 3,* 73–76.

Bundesministerium für Wirtschaft und Energie (BMWi). (2020). *Einsatz des Serviceroboters L2B2 zur Unterstützung der Bürgerdienste der Stadt Ludwigsburg* (KOINNO-Praxisbeispiele). https://www.koinno-bmwk.de/fileadmin/user_upload/praxisbeispiele/KOINNO-Praxisbeispiele_2021_90_Serviceroboter.pdf. Zugegriffen: 15. Okt. 2022.

Czarnecki, C., & Auth, G. (2018). Prozessdigitalisierung durch Robotic Process Automation. In T. Barton, C. Müller, & C. Seel (Hrsg.), *Digitalisierung in Unternehmen: Von den theoretischen Ansätzen zur praktischen Umsetzung* (S. 113–131). Springer Fachmedien. https://doi.org/10.1007/978-3-658-22773-9_7.

Deloitte. (Hrsg.). (2015). *The robots are coming* (Deloitte Insight Report). https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/finance/deloitte-uk-finance-robots-are-coming.pdf. Zugegriffen: 15. Okt. 2022.

Huber, M., & Huber, G. (2011). *Prozess- und Projektmanagement für ITIL®*. Vieweg+Teubner. https://doi.org/10.1007/978-3-8348-8195-3.

Reich, M., & Braasch, T. (2019). Die Revolution der Prozessautomatisierung bei Versicherungsunternehmen: Robotic Process Automation (RPA). In M. Reich & C. Zerres (Hrsg.), *Handbuch Versicherungsmarketing* (S. 291–305). Springer. https://doi.org/10.1007/978-3-662-57755-4_17.

Schaaf, T. (2007). The IT Infrastructure Library (ITIL) – An introduction for practitioners and researchers. *LNCCN, 4543,* 235. https://doi.org/10.1007/978-3-540-72986-0_38.
Smeets, M., Erhard, R. U., & Kaußler, T. (2019). Robotic Process Automation (RPA) in der Finanzwirtschaft: Technologie – Implementierung – Erfolgsfaktoren für Entscheider und Anwender. *Springer Gabler*. https://doi.org/10.1007/978-3-658-26564-9.
Stadt Hamburg. (o. J.). *Kinderleicht zum Kindergeld*. hamburg.de. https://www.hamburg.de/kinderleicht-zum-kindergeld/. Zugegriffen: 11. Apr. 2022.
van der Aalst, W. M. P., Bichler, M., & Heinzl, A. (2018). Robotic process automation. *Business & Information Systems Engineering, 60*(4), 269–272. https://doi.org/10.1007/s12599-018-0542-4.
Wickens, C. D., Hollands, J. G., Banbury, S., & Parasuraman, R. (2016). *Engineering psychology and human performance* (4 Aufl.). Routledge.



Textverarbeitung 10

Zusammenfassung

Dieses Kapitel adressiert KI-Systeme, die sich mit der Verarbeitung, dem Verständnis und der Veränderung von Text in seinen verschiedenen Formen beschäftigen. Ziel ist es, ein besseres Verständnis für diese Art Systeme zu ermöglichen, die Grundlage der eingesetzten Methoden darzulegen und so die Beurteilung der Leistungsfähigkeit textverarbeitender KI zu unterstützen. Dazu werden grundlegende Technologien und Konzepte vorgestellt, sowie ein Einblick in konkrete Anwendungsfälle wie bspw. Einfache Sprache gegeben.

10.1 Einleitung

Vor mehr als viertausend Jahren entstand das erste richtige Alphabet – in Zentralägypten, wo verschiedene Hieroglyphen die Laute der Sprache repräsentierten. Knapp dreitausendfünfhundert Jahre später, 1440, wurde der Buchdruck mit beweglichen Lettern von Gutenberg in Mainz erfunden. Die Geschichte der Menschheit ist von geschriebener Kommunikation geprägt, Kultur wäre ohne sie undenkbar. Daher scheint es naheliegend, dass Texte zu verstehen und zu produzieren auch eine zentrale Aufgabe von KI-Systemen geworden ist.

Die automatisierte Verarbeitung von Textdokumenten stellt gerade für die Verwaltung eine der wichtigsten Neuerung dar, die durch KI erreicht werden können. Aber es ist keine leichte Aufgabe – Ein „nicht“ in einem Satz dreht auf einmal den Sinn des Satzes um. Ein „möglich“ verändert, ob eine Information als gegeben gesichert werden darf oder nicht. Rechtschreibfehler können in jedem Text auftreten – trotzdem sollte und kann man als Mensch verstehen, was in dem Satz gemeint ist. Wir wollen auch effizient sprechen, benutzen Pronomen anstatt

M. Heine et al., *Künstliche Intelligenz in öffentlichen Verwaltungen*, Edition eGov-Campus, https://doi.org/10.1007/978-3-658-40101-6_10

immer die Nomen. Auf welches Wort bezieht sich aber ein Demonstrativpronomen wie „diese" – das ist schon für uns Menschen beim Lesen manchmal schwer. Wie kann eine KI das erreichen?

Dieses Kapitel beschäftigt sich mit der Frage, was KI-Textverarbeitung ist, wie sie funktioniert und wo sie eingesetzt werden kann. Es werden grundlegende Technologien für die Textverarbeitung vorgestellt. Es wird ein spezifisches Beispielprogramm gezeigt, dass versucht, die Semantik in rechtlichen Texten zu erfassen und mit anderen zu vergleichen.

Relevant sind die Rahmenbedingungen, denn die Methoden zur digitalen Textverarbeitung sind inzwischen sehr leistungsfähig. Anfragen, zum Beispiel in Suchmaschinen, werden häufig hervorragend verarbeitet. Gerade bei Texten ist die Frage nach dem Digitalisierungsgrad sehr wichtig. Welche der Dokumente, die einbezogen werden sollen, liegen digital vor? Kommen Dokumente nur per Post? Stehen Texte, mit denen ein System trainiert werden kann, so wie in der Kap. 3 dargestellt, zur Verfügung? KI-Textverarbeitung kann nur gelingen, wenn die Voraussetzungen erfüllt sind.

10.2 Grundlagen der Textverarbeitung

Nachdem nun Fallbeispiele etabliert worden sind, geht es in einem ersten Schritt darum, ein grundlegendes Verständnis dafür zu entwickeln, was „Textverarbeitung" im Kontext von Künstlicher Intelligenz bzw. Machine Learning überhaupt bedeutet. Vereinfacht gesprochen umfasst Textverarbeitung alles, was mit der Kategorisierung von Text zu organisierten Gruppen zu tun hat. Technologien wie Natural Language Processing werden dabei genutzt, um Texte nach bestimmten Bedingungen zu klassifizieren, zu unterteilen oder auszuwerten, sodass neue Erkenntnisse oder Ordnungsstrukturen daraus generiert werden können (vgl. Russel & Norvig, 2012, S. 1001).

Typische Anwendungsfälle, bei denen Textverarbeitung zum Einsatz kommt, sind bspw. Online-Publikationen von Blog- oder News-Artikeln, die automatisiert mit einem thematischen Tag versehen werden. Tonalitätsanalysetools, die den „Ton" eines Zeitungsartikels oder von Konversationen in sozialen Netzwerken bestimmen sollen. Aber auch die Einschätzung von Krisenlagen durch Konversationen über soziale Netzwerke, etwa im Falle einer Naturkatastrophe oder eines Terroranschlags.

Bevor wir nun auf konkrete Techniken eingehen, die bei der Textverarbeitung genutzt werden, muss zunächst dargestellt werden, was bei Prozessen zur Textverarbeitung überhaupt mit dem Text gemacht wird:

- Eine der simpelsten Möglichkeiten ist die Anwendung einfacher statistischer Methoden. So kann die Häufigkeit von bestimmten Worten gezählt und statistisch erfasst werden, um den Text zu kategorisieren. Die Häufigkeit, mit der z. B. ein bestimmter Paragraph oder ein Projektname in einer E-Mail auftaucht, kann so Auskunft darüber geben, wie der Ursprungstext thematisch einzuordnen ist.
- Des Weiteren kann auch das bereits aus Kap. 4 bekannte Pattern Matching genutzt werden, um Texte nach ihren inhärenten Mustern und Strukturen zu klassifizieren. Die Klassifizierung kann beispielsweise darin bestehen, bestimmte Formulierungsfloskeln oder Zitate aus Gesetzestexten zu erkennen. Gleichzeitig können aber auch komplexe Strukturen aufgedeckt werden, die für die jeweiligen Rezipienten ohne die Hilfe des Algorithmus nicht oder nur mit Schwierigkeiten erkennbar gewesen wären. Ein mögliches Beispiel sind Texte, die ein ähnliches Level an komplexer Sprache verwenden o. ä.
- Eine weitere, exemplarische Möglichkeit für die Untersuchung von Texten im Rahmen der Textverarbeitung, die hier vorgestellt werden soll, ist die Möglichkeit, den Text auf Grundlage von verschiedenen, vorher festgelegten Regeln zu untersuchen. Bspw. sollen alle Texte, die Paragraphen enthalten, auf eine bestimmte Art klassifiziert werden oder es werden alle Texte, die eine bestimmte Kombination von Worthäufigkeiten und Merkmalen aufweisen, in eine definierte Gruppe eingeordnet.

Da nun die grundlegenden Prozesse bekannt sind, mit denen Textverarbeitung durchgeführt wird, stellt sich als Nächstes die Frage, auf welche Art und Weise sich unterschiedliche Anwendungen zur Textverarbeitung bei der Umsetzung dieser Methoden unterscheiden. Eine der einfachsten, aber nicht minder relevanten Möglichkeiten, die Fähigkeiten einer Anwendung für die Textverarbeitung einzuschätzen, ist, die Größe der untersuchten Textsegmente zu betrachten, also vereinfach gesagt, wie viel Text kann vom Algorithmus parallel überprüft und verarbeitet werden. Weitere Unterscheidungsmöglichkeiten ergeben sich aus den Fähigkeiten des Algorithmus semantische Unterschiede (Unterschiede auf der Bedeutungsebene des Textes) zu erkennen und zu verarbeiten. Praktische Beispiele für diese semantische Differenzierung sind z. B. die Fähigkeit, Verneinungen als solche zu erkennen und zu bewerten oder aber auch wie gut

der Algorithmus darin ist, die Relevanz verschiedener Textstellen füreinander zu klassifizieren.

Ausgehend von den vorgestellten Methoden und Differenzierungsmöglichkeiten sollen abschließend noch drei Verfahren vorgestellt werden, auf deren Grundlage Algorithmen zur Textverarbeitung operieren können (vgl. Chilakapati, 2019):

Bag of Words
Eine der grundlegendsten Verfahren ist dabei der sogenannte Bag of Words (BoW). Hier wird die Häufigkeit, mit der Worte in einer zu untersuchenden Textmenge auftreten, genutzt, um Classifier zu trainieren, die schlussendlich Aussagen über andere Texte ermöglichen. Stellen wir uns exemplarisch vor, dass wir einen Algorithmus nutzen wollen, um Begrüßungen in E-Mails danach zu klassifizieren, ob sie eher formell oder eher informell von ihrer Ansprache her sind. Der erste Schritt ist das Bereitstellen von Trainingsdaten, mit denen der Algorithmus trainiert werden kann. Sobald wir eine Grundmenge an Beispiel-E-Mails gesammelt haben, müssen diese von uns mit der Dimension, die wir untersuchen wollen, klassifiziert werden. Hier, ob sie formell = 1 oder informell = 0 sind. Diese Trainingsdaten werden dem Algorithmus als Grundlage zur Verfügung gestellt, der, wie bereits beschrieben, schlicht die Häufigkeit der Worte erfasst, die in den Beispieldaten vorkommen. Für unser Beispiel der Sprachformalisierung würde das bedeuten, dass z. B. in 100 als formell beschriebenen E-Mails 95 Mal das Wort „geehrte“ vorkommt, während bei den informellen „Hallo“ die höchste Häufigkeit hat. Auf Grundlage dieser Häufigkeiten und in Abhängigkeit der jeweiligen Zuordnung (formell vs. informell) kann der Algorithmus nun Vorhersagen darüber generieren, ob neue Texte, die ihm zur Verfügung gestellt werden, eher formell oder informell formuliert worden sind.

Word2Vec
Das zweite Verfahren, das betrachtet werden soll, ist die sogenannte Word2Vec (W2V) Methode. Bei W2V geht es darum, aus Worten Vektoren zu errechnen, deren mathematische Nähe zueinander die semantische Nähe der Worte abbildet. Mit einem ausreichend großen Datensatz ermöglichen es die unter dem W2V-Begriff gesammelten Verfahren, in Abhängigkeit des Auftretens der Worte in den Trainingsdaten, recht präzise Aussagen über die Bedeutung eines Wortes zu formulieren. Betrachten wir die Worte Hund, Welpe, Katze und Kätzchen unter der vorgestellten Prämisse, dann würden wir eine größere Nähe zwischen Hund und Welpe bzw. Katze und Kätzchen feststellen, während sich alle in gleicher Distanz zu Mensch befinden könnten. Das Ergebnis von Word2Vec sind nicht nur einzelne Vektoren, sondern ein komplexes Netz, das die Beziehungen der Worte zueinander in Abhängigkeit ihrer

Kosinusähnlichkeit abbildet. Dies ermöglicht es uns, die Worte in Abhängigkeit ihrer Ähnlichkeit zueinander von 0 = keine Ähnlichkeit (90°) bis 1 = volle Ähnlichkeit (0°) darzustellen. Diese Vektorennetze können dann genutzt werden, um weitere Algorithmen mit Informationen zu versorgen. Im Alltag finden wir sie vor allem im Hintergrund von Suchmaschinen oder den diversen E-Commerce Seiten, die wir immer wieder nutzen.

Bidirectional Encoder Representation From Transformers (BERT)
Das letzte Verfahren, das im Rahmen dieses Segments beleuchtet werden soll, ist die Bidirectional Encoder Representation From Transformers oder kurz BERT. BERT ist das modernste der vorgestellten Verfahren und adressiert einige der Schwachstellen zuvor präsentierter Modelle. Die Wichtigsten sind dabei:

1. Die Erkenntnis, dass Bag of Words semantikagnostisch agieren d. h., dass sie nur über die Häufigkeit der zur Verfügung gestellten Worte klassifizieren. Weder bilden sie die Bedeutung des Textes adäquat ab, noch ist gewährleistet, dass „richtige" Worte genutzt wurden, um den Algorithmus anzulernen.
2. Die Tatsache, dass es lediglich einen Vektor pro Wort gibt, was die sinnvolle Einbindung von Dimensionen wie „Kontext, in dem das Wort verwendet worden ist" oder „Position, an dem es im Satz steht" schwierig abbildbar macht. Dies verursacht vor allem dann Probleme, wenn Synonyme/Antonyme die Positionierung des Wortes beeinflussen würden, einem Wort mehrere Bedeutungen zukommen könnte (bspw. um-fahren = ausweichen vs. umfahren = etwas mit dem Vehikel umstoßen/überrollen) oder es sowohl als Verb, aber auch nominalisiert auftreten kann (laufen vs. das Laufen).

BERT versucht dieses Problem aufzulösen oder zumindest zu reduzieren, indem es Wort-Vektoren generiert, die sich abhängig von ihrem aktuellen Ort (in der Word Map) und den anderen Wort-Vektoren in ihrer Umgebung anpassen. Es wird also versucht, sowohl Kontext als auch Semantik mit in die Urteilsfindung des Algorithmus einzubeziehen. Dies geschieht mithilfe eines Aufmerksamkeitsmechanismus: Während zuvor nur einzelne Worte einen Vektor gebildet haben, werden nun die Beziehungen jedes einzelnen Wortes in einem Satz zu allen anderen bekannten Worten als ein sogenannter Kontext-Vektor dargestellt. Hier kommt die „Aufmerksamkeit" ins Spiel. Mit Self-attention, also den Fokus auf ein Wort selbst oder Inter-Attention, den Fokus auf die Beziehung zwischen Worten, wird es möglich, eine Beschreibungskategorie zu entwickeln, mit deren Hilfe Bedeutung verlässlicher abgebildet werden kann. Während wir meistens davon ausgehen, dass ein Satz

nur eine mögliche Auslegung hat und nicht weiter über die einzelnen Worte nachdenken, wird schnell klar, dass einzelne Worte eine engere semantische Beziehung zueinander und somit auch zum gesamten „Wortraum" haben als andere im gleichen Satz.

Im Vergleich mit den anderen vorgestellten Verfahren kann dies als Novum bezeichnet werden, da es nicht nur akkuratere Ergebnisse ermöglicht, sondern auch zum ersten Mal eine Form von Aufmerksamkeit nutzt, die es dem Algorithmus erlaubt, Sätze als multidimensionale Objekte zu verstehen, deren Bedeutung zwar abbildbar ist, sich aber in Abhängigkeit verschiedener Kontexte ständig verändern kann.

Übung

1. Bei welcher Methodik zur Textverarbeitung werden im Wesentlichen die Distanzen einzelner Begriffe genutzt, um ein semantisches Netz zu konstruieren?
   a) Word2Vec
   b) Bag of Words
   c) BERT
   5) Classifier-VII
2. Was bedeutet „semantikagnostisch" im Bezug auf die Textanalyse?
   a) Der Algorithmus arbeitet nur über die Häufigkeit von Begriffen, nicht deren Inhalt
   b) Der Algorithmus kann mit verschiedenen Inhalten arbeiten, nicht nur mit bestimmten
   c) Der Algorithmus lernt die Semantik jedes Textes einzeln und ist nicht vorher schon trainiert
   d) Der Algorithmus beachtet ausschließlich Semantik von Sätzen, nicht aber deren Syntax
3. Welche Stärke hat BERT gegenüber anderen, vektor-basierten Verfahren ?
   a) BERT wurde mit einem deutlich kinderfreundlicheren Datensatz trainiert
   b) BERT wird durch eine unabhängige Expertenkommission kontrolliert
   c) BERT kann mit unterschiedlichen Verwendungen für ein- und dasselbe Wort besser umgehen
   d) BERT erlaubt es die semantische Gesamtbedeutung eines Textes inhärent mit anderen Texten zu vergleichen ohne jedes Wort einzulesen

10.3 Natural-Language-Processing-Bestandteile: Intent, Entity, Kontext und Dialogue Management

Das Verstehen von Sprache ist eine enorm herausfordernde Aufgabe, insbesondere dann, wenn ein System Aussagen oder Aufforderungen im Rahmen eines Dialogs erkennen und korrekt bearbeiten soll. In diesem Kapitel soll es daher etwas detaillierter um die verschiedenen Komponenten einer Nachricht gehen, die z. B. von einem Chat-Bot identifiziert und berücksichtigt werden müssen (vgl. dazu Jain et al., 2018).

Schauen wir uns dafür zunächst verschiedene Beispiele in den beiden Programmen SchreibFix und Memoriali an:

1. Eine Nutzerin fordert SchreibFix auf, ihre Mail am kommenden Tag, um 15.00 Uhr, zu versenden.
2. Ein Sachbearbeiter fragt Memoriali: „Welcher Gutachter wohnt in der Nähe?".
3. Ein Bürger fragt das Memoriali-System, wie die Vermessung von Fenstern funktioniert.
4. Eine Kollegin ist verärgert über SchreibFix und fragt es, warum es so unfähig sei.

Alle diese Ausdrücke werden also an ein System weitergeleitet. Die Bearbeitung von Anfragen, Aussagen und Ausdrücken übernimmt dabei das sogenannte „Dialog Management". Dieses System hat die Aufgabe, die sogenannte „Dialog Handlung" zu identifizieren. Damit ist gemeint, welche Funktion eine bestimmte Aussage im Rahmen des Dialogs erfüllt und welche Antwort diese korrekt bedienen könnte. So stellt „15 km" z. B. keine Antwort für das zweite gegebene Beispiel dar.

Das Dialog Management hat daher die Aufgabe, verschiedene Aspekte des Dialogs und der spezifischen Dialog Handlung zu identifizieren, um eine passende Funktion zuzuschreiben und zu antworten. Dazu gehören im Wesentlichen folgende Aspekte: die Entities einer Aussage bzw. des Dialogs, der Kontext des Dialogs sowie der durch einen Ausdruck dargestellte Intent.

Entities sind dabei die relevanten Elemente bzw. Objekte in der Aussage eines Nutzenden. In Beispiel 3 stellen zum Beispiel die Fenster eine Entity dar, die für die Aussage des Nutzers relevant ist. Die Vermessung stellt ebenso eine Entity dar, die notwendig ist, um den Ausdruck des Satzes zu verstehen. Das System kann diese beiden Entities erkennen und sie in der Bearbeitung der Anfrage einsetzen. Bei der Entwicklung eigener Systeme ist es wichtig, dass klar ist, welche Entities es gibt bzw. welche das System beherrschen soll. Würde in Beispiel 3

nun nicht nach Fenstern, sondern z. B. Türen, gefragt werden, müsste die Entity Tür genauso bekannt sein (würde aber vermutlich zu einem anderen Ergebnis führen). Würde nach z. B. Teppichen gefragt werden, würde Memoriali diese Entity vermutlich nicht bearbeiten können. Es kann daher auch hilfreich sein, den Nutzenden Beispiele für Entities zu zeigen.

Aus der Entity-Übersicht alleine wird jedoch noch keine Dialog Handlung, ansonsten würde „Vermessung Fenster“ ja bereits reichen. Es bleibt aber offen, welche Intention (also welchen Intent) die Person mit diesen Entities verfolgt. Soll eine Vermessung eingereicht werden? Soll eine Vermessung angefordert oder beauftragt werden? In diesem Fall ist der Intent eine Anleitung oder Erklärung. Der Intent wäre also zum Beispiel „Gib mir eine Erklärung“ – die betroffenen Entities wären „Fenster“ und „Vermessung“. Diese Informationen reichen dem Dialog Management, um die Funktion der Aussage zu erkennen und – sofern möglich – eine passende Antwort abzugeben.

Hier zeigt sich auch, wieso eine gebrauchstaugliche Entwicklung solcher Dialog-Systeme nicht immer einfach ist: die Entity „Vermessung“ könnte auch durch „Berechnung“, „Ausmaße“, „Größenangabe“, „Flächenmessung“ oder andere Begriffe adressiert werden. Sind diese im System jedoch nicht hinterlegt, sondern nur der Begriff „Vermessung“ als Entity bekannt, kommt es vermutlich zu einem Fehler. Moderne Sprachassistenten stehen häufig vor diesem Problem und greifen dafür z. B. auf große, semantische Datenbanken zurück.

Jedoch lassen sich auch mit Entity und Intent nicht alle Anfragen zufriedenstellend beantworten. Ein Beispiel dafür stellt die dritte Aussage dar. Um nach der erfolgreich erkannten Funktion der Aussage eine passende Antwort zu liefern, benötigt das Dialog Management nämlich die Information, wo die Frage gestellt worden ist. Dies wird mit dem Kontext des Dialogs ausgedrückt. Alle Meta-Informationen, die für das Dialog Management relevant sind, finden sich hier wieder. Dies kann auch die Identität des Fragestellenden, die Uhrzeit oder die Länge der Unterhaltung beinhalten.

Liegen Kontext, Entities und ein erfolgreich erkannter Intent vor, sind die wichtigsten Bestandteile zur Erkennung einer Anfrage vorhanden.

Übung

1. Eine Nutzerin fragt: „Welche Kosten verursacht eine Neuausstellung des Personalausweises?“ – welche der folgenden Aussagen dazu ist korrekt?
   a) Der Kontext könnte sein: „Neuausstellung Personalausweis„
   b) Bei „Kosten“ handelt es sich nicht um eine Entity
   c) Die Funktion der Frage ist „Kostenauskunft„

   d) Diese Aussage beinhaltet keinen Intent
2. Welche der folgenden Antworten ist kein geeignetes Beispiel für den „Kontext“ einer Anfrage?
   a) Person befindet sich an der Infotheke des Rathauses
   b) Person hat bereits vier fehlgeschlagene Anfragen ausgeführt
   c) Person fragt nach einer Wegbeschreibung
   d) Alle oben genannten sind Kontextbeschreibungen
3. Was ist die Aufgabe des Dialog Managements?
   a) Relevante Elemente der Unterhaltung identifizieren und Antwort auswählen
   b) Überwachung von Antwortlatenz und –tendenz nach Nutzerpräferenz
   c) Überprüfung des Wahrheitsgehaltes von Aussagen vor der Antwort
   d) Aufrechterhaltung personalisierter Informationen zur Präzisierung der Antwort im Text

10.4 Ziele von Textverarbeitung

Da nun die Grundlagen zum Thema der Textverarbeitung etabliert worden sind, kann sich nun den Zielen von Textverarbeitung zugewandt werden. Dazu gilt es zunächst die Anwendungsbereiche klarer zu umreißen.

10.4.1 Anwendungsbereiche von Textverarbeitung

Dabei können wir grundlegende Anwendungsfälle definieren, auf die im weiteren Verlauf genauer eingegangen werden soll:

- **Analyse von Texten:** Hier geht es darum, dass Algorithmen genutzt werden, um Texte zu interpretieren bzw. Informationen aus ihnen zu generieren. Dies kann bspw. in Form einer semantischen Analyse der Fall sein, bei der nach bestimmten Bedeutungen und Inhalten gesucht wird z. B. dem Ton eines Anschreibens oder ob eine bestimmte Klausel im Text enthalten ist. Eine weitere Art der Analyse wäre die Auswertung sonstiger Informationen wie der Häufigkeit bestimmter Worte oder Zeichen.
- **Bearbeitung von Texten:** Bei der Bearbeitung geht es um die Manipulation oder Veränderung von bereits existierenden Textdateien. Anwendungsfälle umfassen die Korrektur von Rechtschreibung und Orthographie in Texten, die Umwandlung in andere Sprachformen (einfache Sprache, Übersetzungen,

etc.) oder aber auch das Schwärzen von Unterlagen in Abhängigkeit ihrer Geheimhaltung.

- **Erstellen von Texten:** Algorithmen können ebenfalls dazu genutzt werden, ganze Texte oder auch einzelne Textsegmente auf Grundlage zuvor definierter Inputs zu generieren. Exemplarisches Beispiel wäre die Generierung des Anschreibens eines Kundenbriefes auf Grundlage von Faktoren wie Ton, Grund des Schreibens (Beispiel SchreibFix).
- **Synthetisierung von Ergebnissen:** Bei der Synthetisierung von Ergebnissen auf Grundlage von Textdateien werden Textverarbeitungsalgorithmen genutzt, um neue Inhalte auf Grundlage des verarbeiteten Textes zu generieren. Dabei ist sie eng mit der Analyse von Texten verbunden. Während „reine" Analyseinhalte meist Metainformationen über den Text widerspiegeln, dieser also noch für das Verständnis des generierten Outputs gebraucht wird, lassen sich die Ergebnisse von Synthetisierung auch unabhängig ihres Ursprungs verstehen.

10.4.2 Fallbeispiele

Nachdem wir diese Klassifizierung von verschiedenen Anwendungsfällen für Textverarbeitung jetzt verinnerlicht haben, sollen nun ein paar anwendungsnahe Beispiele aus dem Arbeitsalltag vorgestellt werden:

Memoriali
Betrachten wir unser Fallbeispiel Memoriali für die Begutachtung von Altbauten. Hier könnte ein Algorithmus zur Textverarbeitung eingesetzt werden, um aus den Massen an Unterlagen, die für die einzelnen Immobilien zur Verfügung gestellt werden, die wichtigsten Argumente zu extrahieren und für den/die Sachbearbeiter/in aufzuarbeiten. Die die Immobilie betreuende Person könnte so beispielsweise nur die Stichworte „Energie" und „Dämmung" in eine Suchmaske eingeben und der Algorithmus würde nicht nur eine Darstellung aller Dokumente liefern, in denen diese Begriffe vorkommen, sondern auch alle aufzeigen, die damit auf einer Bedeutungsebene assoziiert sind (bspw. in Form von Rechnungen für energetische Maßnahmen).

Einfache Sprache
Ein weiteres Beispiel für die Bearbeitung von Text durch Algorithmen, das in unserem Alltag vielleicht häufiger auftaucht als zunächst offensichtlich, ist die Übersetzung von Texten in einfache Sprache. Viele Einrichtungen im Bereich

der öffentlichen Verwaltung bieten die Inhalte ihrer jeweiligen Websites auch in einfacher Sprache an. Einfache oder Bürger-Sprache (vgl. Bürgernnahe Verwaltungssprache, 2002) ist eine vereinfachte Version der Standardsprache, deren Fokus auf der Verständlichkeit des Kommunizierten liegt. Sie zeichnet sich vor allem durch einfachen und kurzen Satzbau sowie die Verwendung von möglichst wenig Fachworten aus. Hier greift die Textverarbeitung ein. Algorithmen können dazu genutzt werden, komplizierte oder fachspezifische Texte zu verarbeiten und eine Version in einfacher Sprache auszugeben. Zusätzlich ist es auch möglich, Texte von Grund auf durch einen Algorithmus in einfacher Sprache schreiben zu lassen wie bspw. dieser Artikel im Guardian, der komplett durch GPT-3 geschrieben wurde: https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3.

SchreibFix

Andere Beispiele für die Erschaffung von ganzen Texten durch einen angelernten Algorithmus ist unser Fallbeispiel SchreibFix. Hier kann das Tool genutzt werden, um auf Grundlage verschiedener Faktoren wie dem gewünschten Ton der E-Mail oder dem bisherigen Verlauf der Konversation, ganze E-Mails zu generieren, die dann durch die Nutzenden sofort verwendet werden können.

Synthetisierte Ergebnisse

Ein weiteres Beispiel ist die Erstellung von synthetisierten Ergebnissen. Diese lassen sich in der Strafverfolgung finden. So können Algorithmen genutzt werden, um die Textdateien, die bei der Beschreibung von gesuchten Personen erstellt werden, zu verarbeiten, sodass aus ihnen heraus neue mediale Formen generiert werden können. Im Falle der Personenbeschreibungen könnte der Algorithmus eingesetzt werden, um aus Informationsteilen wie „Größe“, „Augen- und Haarfarbe“, „Körperbau“ usw. ein Bild zu generieren, das die gesuchte Person zeigt. Ein Prozess, der nicht nur die Erstellung von Fahndungsbildern vereinfacht, sondern es auch ermöglicht, in kurzer Zeit verschiedene Varianten einer gesuchten Person zu generieren.

Übung

1. Was ist kein Ziel von Textverarbeitung?
   a) Die Manipulation bereits existierender Texte
   b) Die Überprüfung von Rechtschreibregeln
   c) Die Analyse des Tonfalls eines Textes
   d) Die Beschreibung eines Bildes durch Text

2. Welche dieser Aussagen ist falsch
   a) Durch Textverarbeitung dürfen sich der Inhalt und Sinn eines Textes nicht ändern
   b) Durch Textverarbeitung können deutlich kürzere oder längere Texte entstehen
   c) Nicht jede Form der Textverarbeitung ist inhärent erklärbar
   d) Die Übersetzung von Texten ist eine spezifische Form der Textverarbeitung
3. Welcher Aspekt ist nicht Teil von einfacher Sprache?
   a) Fokussierung auf simplen Satzbau
   b) Kurze Sätze ohne viele Einschübe/Nebensätze
   c) Nur geringe Verwendung von Fachworten
   d) Große Zeilenabstände für bessere Lesbarkeit

10.5 Fallbeispiel Semantha

In dem vorliegenden Fallbeispiel möchten wir ein existierendes Produkt, welches im Bereich der sprachverarbeitenden KI existiert, genauer unter die Lupe nehmen. Dafür wird zunächst ein Einsatzszenario und die Aufgabe, die das Tool dabei übernehmen kann beschrieben. Im Anschluss wird diskutiert, welche Auswirkungen dies auf Arbeitsprozesse haben kann.

Hinweis: Im vorliegenden Fall geht es um die Nutzung des Tools Semantha von der Firma thingsTHINKING. Die Anwendung der Software wurde für die Diskussion in diesem Buch abgewandelt und entspricht unter Umständen nicht dem (aktuellen) Funktionsumfang des tatsächlichen Produktes.

Cem ist Sachbearbeiter und kümmert sich um alle Anliegen, die mit dem Denkmalschutz in Verbindung stehen. Derzeit ist seine Aufgabe, die Leistungsbeschreibungen eines städtischen Bauherren für Umbaumaßnahmen an einem Gebäude zu prüfen. Er soll feststellen, ob diese mit den Vorschriften übereinstimmen. Dieser Prozess ist mühselig, da das Auftragsbuch enorm umfangreich ist und für verschiedene Etagen, Fenstergrößen, etc. einzelne Beschreibungen des Austauschprozesses angefertigt worden sind. Cem entscheidet, dass eine genaue Prüfung zu viel Zeit in Anspruch nehmen würde. Einerseits hat er selbst genug zu tun, andererseits ist das Projekt schon im Verzug. Er wählt die Leistungsbeschreibung von einem Fenster je Etage und segnet danach die Leistungsübersicht ab.

Mithilfe eines sprachverarbeitenden Tools könnte Cem hier auch anders vorgehen. Hinter Semantha steckt ein Algorithmus, der in der Lage ist, den semantischen Inhalt von Texten zu erfassen, auch wenn diese nicht wortwörtlich

miteinander übereinstimmen. Dies kann hilfreich sein, wenn zum Beispiel unterschiedliche Rechtstexte schnell miteinander verglichen werden müssen. Auch bei der Analyse oder dem Vergleich von Verträgen können Tools zur Verarbeitung semantischer Daten genutzt werden.

Was würde in dem Beispiel genau geschehen? Anstatt den Auftragstext komplett zu lesen, hat Cem sich entschieden, die Korrektheit stichprobenartig zu überprüfen. Er versucht so, seine Arbeitslast zu reduzieren und dennoch ein verlässliches Ergebnis zu erreichen. Die Stichproben wählt er zufällig, hier z. B. in jeder Etage ein Fenster, aus.

Beim Einsatz von Semantha würde sich dieser Arbeitsprozess verändern. Die Sprachverarbeitungssoftware würde jede Leistungsbeschreibung automatisiert mit den entsprechenden Vorschriften für den Denkmalschutz bzw. entsprechende Umbaumaßnahmen abgleichen. Die Stärke eines KI-basierten Ansatzes ist dabei, dass zum Beispiel ein Begriff wie „nicht abschließend" und „undicht" im Bezug auf Fenster als gleichwertig betrachtet werden. Dadurch ist es möglich, Texte miteinander zu vergleichen, die zwar unterschiedliche Autoren bzw. Wortschätze haben, aber dasselbe Thema behandeln.

In unserem Beispiel könnte Semantha so – deutlich schneller als Cem – die gesamten Vertragsunterlagen prüfen. Dabei werden auffällige Stellen markiert, z. B. wenn in der Vorschrift „gleichwertiges Erscheinungsbild" und im Vertrag „moderne Ästhetik" stehen. Diese Punkte kann Cem dann in Ruhe prüfen und entscheiden, ob die von der KI angemerkten Stellen tatsächlich eine Abweichung darstellen oder nicht. Fälle, bei denen das System sich unsicher ist oder Fälle, bei denen es eine klare Abweichung vermutet, können priorisiert bearbeitet werden.

In diesem Beispiel übernimmt die das textverarbeitende KI-System also einen Teil der menschlichen Aufgabe – die Sortierung der Stellen, die zu prüfen sind. Besonders erfahrene Sachbearbeitende können hier vielleicht auf Erfahrung zurückgreifen; ggf. gibt es auch interne Richtlinien, welche Stellen besonders geprüft werden müssen. Aber gerade in Fällen, wo die Masse an Text sehr groß ist, kann diese Priorisierung sehr schwierig sein und erfordert entweder viel Zeit oder muss zufällig vorgenommen werden.

Textverarbeitende KI wie Sematha kann die Verwaltung hierbei unterstützen, indem im Rahmen einer semantischen Analyse große Mengen an Daten untersucht und für die Bearbeitung durch menschliche Sachbearbeitende priorisiert werden.

10.6 Beispiele einfache Sprache

Im Rahmen dieses Fallbeispiels wollen wir uns noch einmal etwas detaillierter mit dem Konzept der „Einfachen Sprache“ im Verwaltungskontext und den möglichen Anwendungsbereichen von Algorithmen in Bezug auf diese beschäftigen.

Wie bereits zu den Zielen von Textverarbeitung erläutert, handelt es sich bei Einfacher Sprache um eine simplifizierte Variante unserer Alltagssprache. Sie wurde entwickelt, um eine möglichst bürgernahe und direkte Kommunikation zu ermöglichen und verzichtet dabei unter anderem auf komplexe Satzstrukturen und unnötige Fachbegriffe.

Die Zielgruppe der einfachen Sprache sind vor allem Menschen mit geringer sprachlicher Kompetenz, Personen mit geistigen Behinderungen oder Menschen, die Deutsch als Fremdsprache gelernt, aber noch kein hohes Sprachniveau erlangt haben. Vereinfacht gesagt geht es darum, die Behörden vom sperrigen Amtsdeutsch zu befreien und eine Kommunikationsform zu wählen, die möglichst zugänglich für alle Beteiligten ist. Einfache Sprache ist dabei vom Konzept der Leichten Sprache abzugrenzen, bei der es sich um eine nach festen Regeln agierende Sprachvariante handelt, die in den 1970ern explizit aus der Perspektive der Barrierefreiheit und Inklusion entwickelt worden ist (vgl. Kellermann, 2014).

Bei der Produktion und Konzeption von Texten in Einfacher Sprache kann der Einsatz von Algorithmen zur Textverarbeitung unterstützen. Wie zuvor bereits beschrieben wurde, können diese Algorithmen so trainiert werden, dass sie bereits existierende Textdateien zu neuen Texten umformen oder aber neue Texte aus Stichworten und Anmerkungen generieren. Dies ist genau der Anwendungsfall, der in Bezug auf Einfache Sprache benötigt wird. So können bereits bestehende Texte in komplexeren Sprachformen wie z. B. Amtsdeutsch in den mit Einfacher Sprache angelernten Algorithmus eingegeben werden, sodass dieser aus ihnen Texte mit den gleichen Informationen, aber einfacheren Satzstrukturen oder ähnlichen Merkmalen der Einfachen Sprache bildet. Dies ermöglicht es, bereits bestehende Informationen, die schon auf den Webseiten der Behörden existieren, anzupassen, ohne dass jeder einzelne Text händisch neu geschrieben werden muss, was einen großen Kosten- und Arbeitsaufwand mit sich bringen würde.

Ein weiterer Anwendungsfall, bei dem KI-Systeme unterstützen können, ist die Neuerstellung von Inhalten. Hier können Sprachtools wie Capito eingesetzt werden, die schon während des Schreibprozesses Informationen über den Komplexitätsgrad des Textes geben und es den Autor/innen so ermöglichen, sich an ihre Zielgruppen anzupassen. Die Einschätzung der Sprachkomplexität ist von

großer Bedeutung, wenn es beispielsweise darum geht, Texte mit wichtigen Informationen, wie das richtige Verhalten in Notfällen, für alle Adressaten zugänglich zu machen.

Grundsätzlich kann Einfache Sprache an allen Orten eingesetzt werden, an denen die Kommunikation mit einer sehr divers sprachkompetenten Gruppe von Personen gewährleistet werden muss. Je wichtiger die zu kommunizierenden Informationen sind, umso mehr sollte Einfache Sprache einbezogen werden. Es ist daher wenig überraschend, dass die Haupteinsatzgebiete vor allem auf den Webseiten und Informationsbroschüren von Behörden und Ämtern sowie medizinischen Einrichtungen zu finden sind.

Probleme von Einfacher Sprache bzw. deren technischer Einbindung durch Textverarbeitung sind vor allem die uneinheitlichen Regeln. Anders als bei Leichter Sprache gibt es keine zentral bestimmten Vorgaben für die Schreibweise. Vielmehr haben verschiedene Institutionen Richtlinien, die für Einfache Sprache genutzt werden können. Ein damit eng verbundener Problempunkt ist die zuvor vorgestellte Güte der Daten. Die Faktoren, nach denen ein lernender Algorithmus schlussendlich beurteilt, sind abhängig von den Datensätzen, mit denen er angelernt wird. Fehlerhafte oder von den gewünschten Regeln abweichende Datensätze könnten also zu unerwünschten bzw. falschen Ergebnissen führen.

Die Algorithmen, die für diese Art von Textverarbeitung angelernt werden, bedienen sich dabei der bereits in Abschn. „10.2 Grundlagen der Textverarbeitung“ vorgestellten Techniken zur Textverarbeitung. Grundlage bietet dabei vor allem die Entwicklung von Algorithmen auf Basis bestimmter Regelstrukturen. Beispielsweise kann ein Algorithmus, der lediglich aus der Anzahl der Worte pro Satz und der Länge eben dieser Worte einen Kennwert bildet, genutzt werden, um Aussagen über die Verständlichkeit des zu untersuchenden Textes zu treffen.

Eine weitere Möglichkeit ist der Einsatz von Deep Learning. Hier werden dem Algorithmus Trainingsdaten in großen Mengen zur Verfügung gestellt, auf deren Grundlage er trainiert wird und lernt, bestimmte Muster zu erkennen und Entscheidungen zu treffen. So könnte man bspw. großen Mengen an Texten in Einfacher Sprache zur Verfügung stellen, um dann schlussendlich neue Texte nach dem Konzept der Einfachheit bewerten zu lassen und so Stufen der Sprachkomplexität zu entwickeln. Diese Komplexitätseinschätzung durch den Algorithmus hilft dann den Nutzer/innen dabei, ihre neuen Textstücke an ihre Zielgruppen anzupassen. Auch die bereits vorgestellte Idee eines Übersetzungstools, bei dem normale Sprache in ein Feld eingegeben und Einfache Sprache ausgegeben wird, ließe sich mit dieser Art Algorithmen realisieren.

10.7 Aufgaben zum eigenen Anwendungsfall

In vielen Anwendungsfällen innerhalb der Verwaltung wird die Nutzung von textverarbeitenden Systemen eine wichtige Rolle spielen. Unterschiedliche Technologien sind dabei für bestimmte Aufgaben gerüstet – für andere nicht.

- Erörtern Sie, welche Schritte notwendig und zu prüfen sind, bevor Sie den Einsatz von textverarbeitender KI in Angriff nehmen. Bewerten Sie dabei einerseits den aktuellen Zustand aber auch den Arbeitsaufwand sowie etwaige Erleichterungen zur Integration einer textverarbeitenden KI in Ihrem Projekt.
- Diskutieren Sie, welche der vorgestellten Technologien eingesetzt werden können. Wählen Sie mindestens zwei Technologien aus und stellen Sie die Unterschiede anhand eines Anwendungsfalles dar. Wählen Sie dabei ein Beispiel aus Ihrem Projekt oder gehen Sie davon aus, dass mittels eines Chatbots Fragen von externen Partnern oder Bürger/innen beantwortet werden können sollen.
- Erstellen Sie für Ihren Anwendungsfall drei Beispielfragen und schildern Sie, was Intent, Entity & Kontext in diesen Fällen sein könnten.

Literatur

Bürgernahe Verwaltungssprache. (2002). https://www.bva.bund.de/SharedDocs/Downloads/DE/Oeffentlichkeitsarbeit/Buergernahe_Verwaltungssprache_BBB.pdf?__blob=publicationFile&v=5. Zugegriffen: 15. Okt. 2022.

Chilakapati, A. (23. September 2019). BoW to BERT. Data exploration. https://xplordat.com/2019/09/23/bow-to-bert/. Zugegriffen: 15. Okt. 2022.

Jain, M., Kota, R., Kumar, P., & Patel, S. N. (2018). Convey: Exploring the use of a context view for chatbots. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (S. 1–6). Association for Computing Machinery. https://doi.org/10.1145/3173574.3174042.

Kellermann, G. (2014). Leichte und Einfache Sprache – Versuch einer Definition | APuZ. bpb.de, https://www.bpb.de/apuz/179341/leichte-und-einfache-sprache-versuch-einer-definition. Zugegriffen: 15. Okt. 2022.

Russell, S., & Norvig, P. (2012). *Künstliche Intelligenz* (Bd. 2). Pearson Studium.

Weiterführende Literatur und Online-Artikel

Alammar, J. (2018). The illustrated transformer, von http://jalammar.github.io/illustrated-transformer/. Zugegriffen: 15. Okt. 2022.

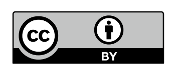
Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. https://arxiv.org/abs/1810.04805v2. Zugegriffen: 15. Okt. 2022.

Roth, U. (2020). Künstliche Intelligenz schreibt in Einfacher Sprache/Plain Language, https://leichtgesagt.eu/kuenstliche-intelligenz-schreibt-in-einfacher-sprache-plain-language. Zugegriffen: 15. Okt. 2022.

McTear, M. F., Callejas, Z., & Griol, D. (2016). The conversational interface (Bd. 6, No. 94, S. 102). Springer.

Quamar, A., Lei, C., Miller, D., Ozcan, F., Kreulen, J., Moore, R. J., & Efthymiou, V. (2020). An ontology-based conversation system for knowledge bases. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data,* 361–376. https://doi.org/10.1145/3318464.3386139.

Zhang, Y., Jin, R., & Zhou, Z. H. (2010). Understanding bag-of-words model: A statistical framework. *International journal of machine learning and cybernetics, 1,* 43–52.





KI & Ethik

11

Zusammenfassung

In diesem Kapitel wird die Rolle der Ethik bei KI-Anwendungen betrachtet. Dafür werden zunächst grundlegende Begriffe definiert und ethische Aspekte beim Einsatz von KI-Systemen vorgestellt. Es werden Fragen zur Verfügung gestellt, die für die ethische Bewertung von KI-Anwendungen herangezogen werden können. Um die Auseinandersetzung mit ethischen Fragen plastischer darzustellen und eine weitere Perspektive zu bieten, endet das Kapitel mit einem Interview, das mit einem Wissenschaftler aus dem Bereich Angewandte Ethik geführt wurde.

11.1 Einleitung

Wenn KI-Systeme in Entscheidungsprozesse involviert sind oder in anderer Weise unser Leben beeinflussen, dann muss dies auf ethische Weise passieren. Das betrifft nicht nur die konkrete Entscheidung selbst, sondern auch z. B. was dieser zugrunde liegt (z. B. die Datenbasis). Doch was ist mit ethischem Verhalten gemeint, speziell im Bereich KI? Zur Beantwortung dieser Frage werden zuerst drei Fallbeispiele vorgestellt (11.2), dann wird Ethik definiert (11.3), die ethischen Aspekte beim Einsatz von KI-Systemen betrachtet (11.4) und Fragen zur ethischen Bewertung von KI-Anwendungen vorgestellt (11.5). Anschließend werden kurz die zentralen Ergebnisse eines Interviews zu ethischen Aspekten von KI zusammengefasst (11.6). Abschließend folgen Fragen zur Bewertung von KI-Anwendungen (11.7), Übungsfragen (11.8) sowie Aufgaben zum eigenen Anwendungsfall (11.9).

M. Heine et al., *Künstliche Intelligenz in öffentlichen Verwaltungen*, Edition eGov-Campus, https://doi.org/10.1007/978-3-658-40101-6_11

11.2 Fallbeispiele

Betrachtet man die Entscheidungen von KI-Anwendungen und inwieweit Menschen sie als ethisch bezeichnen würden, lassen sich gute, schlechte, und hässliche Anwendungen identifizieren.

Gute Anwendungen können Verzerrungen, insbesondere Vorurteile, in Entscheidungen reduzieren. Ein typisches Beispiel sind Personalentscheidungen, wenn z. B. die verantwortliche Person bestimmte Gruppen unangemessen bevorzugt. Zwar ist die Vorstellung, dass eine KI-Anwendung objektiv und damit vorurteilsfrei ist, sehr gewagt, eine KI-Anwendung kann allerdings Entscheidungen transparenter machen (siehe auch Unterkapitel 7.5). Durch diese Transparenz können Entscheidungen auf mögliche irrelevante Verzerrungen untersucht und gegebenenfalls reduziert werden.

Schlechte Anwendungen können Entscheidungen treffen, die vielleicht rational sind, viele Menschen aber für ethisch falsch halten. Ein fiktionales Beispiel sieht man im Film „I, Robot". Ein Roboter entscheidet, welche von zwei Personen die größten Überlebenschancen hat und fällt auf dieser Basis die Entscheidung: das System rettet den erwachsenen Mann und lässt das kleine Mädchen ertrinken. Viele Personen würden dem – vermutlich aus gutem Grund – nicht zustimmen.

Hässliche Anwendungen verändern die Situation sehr subtil. Bei der Verwendung von Social-Media-Anwendungen werden Nutzer heute schon von Algorithmen beeinflusst, oft ohne es zu merken. Unsere Welt verändert sich unmerklich, unser Verhalten wird so manipuliert, dass Nutzer die Anwendung möglichst lange verwenden. Diese Manipulation kann auch ohne negative Absicht auftreten.

Die Frage ist jetzt, wie muss eine KI-Anwendung gestaltet sein, damit diese sich möglichst „ethisch" verhält. Worauf muss man dabei achten?

11.3 Ethik und KI – sieben Thesen

Was ist mit Ethik gemeint? Für die Betrachtung von KI und Ethik ist zentral, dass die Ethik als philosophische Disziplin die grundsätzliche Frage behandelt: Was ist gut/richtig vs. was ist schlecht/falsch? Umgesetzt ist sie dann zum einen im Recht, aber auch in ungeschriebenen Idealvorstellungen. Oder näher an Ethik und KI formuliert (Markkula Center for Applied Ethics, zitiert via IBM, 2022): „Ethics is based on well-founded standards of right and wrong that prescribe what humans ought to do, usually in terms of rights, obligations, benefits to society, fairness, or specific virtues." bzw. übersetzt: „Ethik basiert auf wohlbegründeten

Standards von Richtig und Falsch, die vorschreiben, was Menschen machen sollten, normalerweise in Begriffen von Rechten, Verpflichtungen, Vorteilen für die Gesellschaft, Fairness oder spezifischen Werten."

Warum ist Ethik bei KI besonders wichtig? Die folgenden Abschnitte erläutern sieben Thesen zu dieser Frage.

11.3.1 Menschen würden von einer KI ethisches Verhalten erwarten

Wenn eine KI Empfehlungen gibt oder Entscheidungen trifft, dann würden die Nutzer ethisches Verhalten erwarten. Unethische Vorschläge wären bei Entscheidungen, für die man sich rechtfertigen muss (inkl. vor sich selbst), wenig hilfreich.

11.3.2 Der Einsatz von KI macht unsere Entscheidungsregeln und die Konsequenzen transparent

Der Einsatz von KI macht unsere Entscheidungen, und damit die Grundlage und die Konsequenzen, transparent. Die Entscheidungsprozesse und deren Ergebnisse werden bei KI leicht beobachtbar und diskutierbar. Menschen müssen entweder über explizite Regeln oder (beim maschinellen Lernen) über Rückmeldung, welches Verhalten richtig ist, explizit machen, was das – auch ethisch – richtige Verhalten ist. Gerade diese Transparenz, wie ist die Entscheidung zustande gekommen und welche Auswirkungen hat es, macht die Werte-Frage explizit. Es wirft die Frage auf: Was ist eine gute, faire, korrekte Entscheidung? Diese Frage kann nicht nur anhand von Regeln oder Vorgaben beantwortet werden, sondern auch mit Blick auf die Anwendung über viele Fälle hinweg. Man kann sich z. B. alle Anträge der letzten zehn Jahre ansehen und relativ leicht überprüfen, ob Entscheidungen z. B. bestimmte Gruppen ungerechtfertigt benachteiligen.

11.3.3 Der Einsatz von KI erfordert die Festlegung auf soziale und moralische Normen

Wenn man die Regeln oder die Richtigkeit von KI-Entscheidungen bewertet, muss man genau überlegen, an welchen Normen der Gesellschaft man sich orientiert. Die IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems

(zitiert via IBM, 2022) formuliert dazu: „If machines engage in human communities as autonomous agents, then those agents will be expected to follow the community's social and moral norms. A necessary step in enabling machines to do so is to identify these norms. But whose norms?" übersetzt: „Wenn eine Maschine in menschlichen Gemeinschaften als autonome Handelnde eingreifen, dann wird man von ihnen erwarten, dass sie den sozialen und moralischen Normen der Gemeinschaft folgen. Ein notwendiger Schritt darin, die Maschinen in die Lage zu versetzen, dies zu machen, ist diese Normen zu identifizieren. Aber wessen Normen?"

Der Einsatz von KI erfordert es, sich darüber Gedanken zu machen, auf welcher Grundlage man Entscheidungen trifft, was die Konsequenzen sind, und letztendlich, welche Welt oder welches Weltbild vertreten wird.

11.3.4 Logik und Rationalität „der KI" ist ein Trugschluss und wäre auch nicht wünschenswert

Die (oft nur angenommene) Logik und Rationalität von KI ist ein Trugschluss. Es gilt auch hier Garbage in, Garbage out (GIGO, übersetzt als „Müll rein, Müll raus") – wenn z. B. in den Trainingsdaten Verzerrungen enthalten sind, weil die Personen, welche die Entscheidungen getroffen haben, auch nicht vorurteilsfrei waren, dann wird das trainierte System diese Verzerrungen auch aufweisen (vgl. Abschn. 11.4.2).

Des Weiteren ist der Mensch kein „homo oeconomicus", der Entscheidungen rein rational und auf Basis der verfügbaren Informationen trifft. Menschen haben schlicht nicht die Ressourcen, um bei allen Entscheidungen rational zu handeln. Eine rein logische Welt wäre auch „unmenschlich". Nicht ohne Grund gibt es auch bei (einigen) Entscheidungen in der öffentlichen Verwaltung einen Ermessensspielraum. Dieser lässt sich aber nur schwer in Regeln gießen.

11.3.5 Wissenschaft und Technik sind wertneutral

Ein weiterer Punkt ist, dass KI – ebenso wie jegliche Wissenschaft und Technik – nur ein Werkzeug ist, das wertneutral ist. Die Wissenschaft und Technik besitzt keine inhärenten Sicherungsmechanismen, die einen unethischen Einsatz verhindern würden. Richard Dawkins beschreibt dies wie folgt: „Scientific and technological progress themselves are value-neutral. They are just very good at doing what they do. If you want to do selfish, greedy, intolerant and violent

things, scientific technology will provide you with by far the most efficient way of doing so. But if you want to do good, to solve the world's problems, to progress in the best value-laden sense, once again, there is no better means to those ends than the scientific way.“ übersetzt (mittels KI, der kostenlosen Version von DeepL.com/Translator): „Der wissenschaftliche und technische Fortschritt selbst ist wertneutral. Sie sind nur sehr gut darin, das zu tun, was sie tun. Wenn Sie egoistische, gierige, intolerante und gewalttätige Dinge tun wollen, bietet Ihnen die wissenschaftliche Technologie die bei weitem effizienteste Möglichkeit, dies zu tun. Wenn Sie aber Gutes tun, die Probleme der Welt lösen und im besten werteorientierten Sinne fortschreiten wollen, dann gibt es wiederum kein besseres Mittel als die Wissenschaft, um diese Ziele zu erreichen.“

Technik ist wertneutral – es kann für jeden Zweck eingesetzt werden, ob gut oder schlecht. Man kann z. B. Microsoft Excel verwenden, um die sozialen Mittel besser zu verwalten, man könnte damit aber auch Verbrechen planen und organisieren. Die Wertneutralität von Technik hat bei KI aber eine besondere Bedeutung bekommen, weil KI-Systeme auch autonom agieren können. Je nach Automationsgrad fällt ein System Entscheidungen oder liefert zumindest Hilfe bei der Entscheidungsfindung. Die Bewertung inwiefern das Verhalten von KI-Systemen ethisch ist, muss dabei vom Menschen kommen. Es gibt allerdings Versuche, KI für die Bewertung von ethischen Fragen einzusetzen – z. B. delphi.allenai.org. Die Ergebnisse sind allerdings mit Vorsicht zu behandeln.

11.3.6 Computer können einen extrem starken Einfluss auf unser Verhalten haben

Dass beim Einsatz von KI unweigerlich Computer verwendet werden, kann ein weiteres ethisches Problem darstellen. Computer haben viele Stärken, wenn es um die Unterstützung von Menschen geht, insbesondere bei der Verhaltensänderung (z. B. über Persuasive Technology). Beispiele sind Mindfulness-Apps oder auch Fitness-Tracker, welche die Nutzer zu selbst gewähltem, besserem Verhalten motivieren sollen. Fogg (2003) nennt als Vorteile von Computern unter anderem deren Hartnäckigkeit, Anonymität, Umgang mit großen Datenmengen, Nutzung unterschiedlicher Modalitäten, leichte Skalierbarkeit (da sich Anwendungen leicht kopieren und verbreiten lassen) und Einsetzbarkeit an Orten, an denen andere Menschen nicht willkommen sind. Entsprechend können Computer sehr gut darin sein, Menschen bei der Änderung von Gewohnheiten zu unterstützen.

All diese Stärken werden zu einem Problem, wenn ein KI-System – aus unserer Sicht – unethische Entscheidungen trifft. Es ist ein beständiger Einfluss oder auch Druck, der immer und immer wieder ausgeübt wird.

11.3.7 Die Verantwortung liegt beim Menschen

All diese Punkte sorgen dafür, dass man am Thema Ethik nicht vorbeikommt. Und das ist nichts, was man an KI-Systeme abgeben kann. Die Verantwortung kann nur der Mensch übernehmen – schließlich kann man schlecht sagen: „Ich habe nur die Anweisungen der KI befolgt.“. Entsprechend muss sichergestellt werden, dass ein KI-System ethisch operiert. Nur was gehört bei einem KI-System zum ethischen Verhalten dazu – und was sind mögliche Risiken?

11.4 Ethische Aspekte beim Einsatz von KI-Systemen

Was sind ethische Aspekte beim Einsatz von KI? Zunächst einmal, Ethik geht über die konkrete Empfehlung oder Entscheidung hinaus. Es umfasst auch Aspekte, die vor der Empfehlung bzw. Entscheidung passieren sowie die Konsequenzen (inkl. langfristiger Auswirkungen). Darunter fällt (u. a. Wing, 2021) zum Beispiel die Datenbasis (Wurden die Daten auf ethische Weise gesammelt?), der Betrieb des Systems (Wie ist die Kosten-Nutzen-Rechnung bezüglich Energieverbrauch/Nachhaltigkeit?), das Urteilen/Entscheiden des KI-Systems (Werden unverzerrte, faire Entscheidungen getroffen?) und die Verwendung des KI-Systems (Werden die Ergebnisse auf ethische Weise genutzt?).

Hauptaspekte bei KI und Ethik sind vor allem „Fairness“ (Abschn. 11.4.1), die Vermeidung von Verzerrungen (biases, Abschn. 11.4.2), Datenschutz & Privatsphäre (Abschn. 11.4.3), Vermeidung von (oft subtilen) Beeinflussungen (Abschn. 11.4.4), die Auswirkungen von KI auf die öffentliche Verwaltung (Abschn. 11.4.5), sowie gesellschaftliche Auswirkungen (Abschn. 11.4.6). Außerdem ergeben sich bei der ethischen Betrachtung oft Zielkonflikte mit anderen Kriterien.

11.4.1 Fairness

Fairness ist ein überraschend komplexer Begriff. Im Kern betrifft es das Vermeiden oder Verhindern von unerwünschter bzw. ungerechtfertigter Diskriminierung. Aber wie Fairness konkret definiert wird, ist alles andere als eindeutig.

Fairness ist auch das Thema, das bei KI viel Aufmerksamkeit bekommen hat. U. a. weil niemand aufgrund von irrelevanten Eigenschaften schlechter behandelt werden soll und eine solche Behandlung – die KI sichtbar macht (vgl. Abschn. 11.2) – üblicherweise eine hohe öffentliche bzw. mediale Aufmerksamkeit auslöst. Entsprechend sind die großen Unternehmen dabei, Fairness explizit zu adressieren, z. B. IBM's AI Fairness, Google's TensorFlow kit, Microsoft's Fairlearn, Facebook's Fairness Flow oder Amazon & NSF's Fairness in AI (Wing, 2021).

Fairness ist insbesondere in der öffentlichen Verwaltung auch rechtlich relevant, denn es gibt das „… Verbot, gleiche soziale Sachverhalte ungleich oder ungleiche gleich zu behandeln, es sei denn, ein abweichendes Vorgehen wäre sachlich gerechtfertigt." (Poretschkin et al., 2021). Entsprechend darf keine Benachteiligung aufgrund irrelevanter Kriterien erfolgen. Dies *kann* z. B. Nationalität, ethnische Herkunft, Geschlecht, Religion/Weltanschauung, Behinderung, Altersgruppe oder sexuelle Identität beinhalten – sofern diese irrelevant für die entsprechende Entscheidung sind (Poretschkin et al., 2021). Fairness ist vor allem relevant bei Systemen, die Vorhersagen machen (wie z. B. bei vielen Decision Support Systems).

Wie kann man Fairness beurteilen?

Eine vermeintlich einfache Anforderung wie Fairness ist überraschend komplex, weil man Fairness nach verschiedenen Standards mit verschiedenen Werten beurteilen kann (vgl. Verma und Rubin, 2018).

Auf Einzelfall-Ebene ist eine richtige Entscheidung recht einfach. Das KI-System kann die richtige Entscheidung treffen (richtig positiv – Antrag wird bewilligt, Person ist auch berechtigt bzw. richtig negativ – Antrag wird nicht bewilligt, Person ist auch nicht berechtigt). Sie kann aber auch eine falsche Entscheidung treffen (falsch positiv – Antrag wird bewilligt, Person ist aber nicht berechtigt bzw. falsch negativ – Antrag wird nicht bewilligt, Person ist aber berechtigt).

Schaut man sich jetzt die Qualität von vielen Vorhersagen an, kann man unterschiedliche Qualitätsmaße berechnen. Zum Beispiel die Wahrscheinlichkeit, dass eine als berechtigt vorhergesagte Person auch in Wirklichkeit berechtigt ist, oder eine berechtigte Person auch richtig als berechtigt vorhergesagt wird, oder eine als berechtigt vorhergesagte Person in Wirklichkeit nicht berechtigt ist, oder eine

unberechtigte Person fälschlicherweise als berechtigt vorhergesagt wird, oder eine als nicht berechtigt vorhergesagte Person in Wirklichkeit berechtigt ist, oder eine berechtigte Person fälschlicherweise als unberechtigt vorhergesagt wird, oder eine als nicht berechtigt vorhergesagte Person in Wirklichkeit auch nicht berechtigt ist, oder eine unberechtigte Person auch richtig als unberechtigt vorhergesagt wird (siehe Verma & Rubin, 2018). Was hat das mit Fairness zu tun?

Mit diesen Berechnungen kann man unterschiedliche Maße für die Vorhersagequalität für unterschiedliche Gruppen berechnen und vergleichen (vgl. Verma & Rubin, 2018). Je nachdem welche Werte zugrunde liegen, kann man zum Ergebnis kommen, dass ein Algorithmus fair ist oder nicht. Bei „group fairness" ist die Wahrscheinlichkeit, den Antrag bewilligt zu bekommen, die gleiche für unterschiedliche Gruppen. Die „conditional statistical parity" ist ähnlich wie die „group fairness", aber legitime Faktoren dürfen mit einbezogen werden. Bei der „predictive parity" ist die Wahrscheinlichkeit, dass z. B. ein Antrag korrekt bewilligt wird, die Gleiche für unterschiedliche Gruppen.

Bei der Fairness geht es oft um Gruppenfairness vs. individuelle Fairness. Das sind zwei sehr unterschiedliche Konzepte. Bei der Gruppenfairness sind die Ergebnisse der Anwendung für alle vorhandenen Gruppen vergleichbar (gleiche Verteilung auf verschiedene Gruppen, gleiche Vorhersagequalität in allen Gruppen). Bei der individuellen Fairness werden gleiche Individuen gleich behandelt. Dies berührt das Spannungsfeld von „Ergebnisgleichheit" vs. Chancengleichheit.

Diese und andere Beispiele (siehe dazu Verma & Rubin, 2018) zeigen, wie komplex ein intuitiv einfaches Konzept wie „Fairness" plötzlich sein kann. Das war es vorher auch schon, aber durch die Datenbasis, die Möglichkeit, den Entscheidungsprozess, die Ergebnisse und die Konsequenzen klar vor sich zu haben, wird es plötzlich – wie eingangs geschrieben – beobachtbar und diskutierbar. Die Transparenz macht die Werte-Frage – was ist eine gute, faire, korrekte Entscheidung – explizit.

11.4.2 Vermeidung von Verzerrungen (biases)

Insbesondere wenn Fairness-Kriterien verletzt sind, stellt sich die Frage – woher kommen diese Abweichungen? Dies führt zum Thema der Verzerrungen (engl. biases). Es kann sein, dass während der Entwicklung eines KI-Systems Verzerrungen eingeflossen sind, die dort eigentlich nicht vorkommen sollten.

Diese Verzerrungen können leicht einfließen, da an der Entwicklung eines KI-Systems – v. a. wenn maschinelles Lernen verwendet wird – viele Personen

beteiligt sind. Dabei unterscheidet man zwischen Verzerrungen bei der Erstellung der Trainingsdaten (data-creation bias), Verzerrungen bei der Problemdefinition, Verzerrungen bezogen auf Algorithmen und Datenanalyse und Verzerrungen bezogen auf die Evaluation und Validierung durch Menschen (siehe Srinivasan & Chander, 2021). Diese Verzerrungen können sich über die Entwicklungspipeline ansammeln. Von der Erstellung der Trainingsdaten, der Definition des Problems, den Algorithmen und der Datenanalyse, zur Evaluation und Validierung durch Menschen.

Insbesondere die Definition des Problems ist interessant. Betrachtet man z. B. die Agentur für Arbeit, dann stellt sich die Frage, wie eine gute Stellenvermittlung so messbar gemacht wird, dass man dem KI-System Rückmeldung geben kann, was eine richtige oder falsche Entscheidung war (Operationalisierung). Man kann als Ziel haben, dass möglichst wenige Personen zu einem gegebenen Zeitpunkt arbeitssuchend sind, d. h. Ziel ist, möglichst schnell eine ausreichend passende Stelle finden. Man kann aber auch als Ziel haben, dass Personen eine Stelle finden, die langfristig zu ihnen passt und in der sie sich auch weiter entwickeln können. Das ist eine Frage, wie man das Problem framed (also rahmt), was auch bestimmt, welche Daten ausgewählt werden (und welche Entscheidungen den Daten zugrunde liegen), um das Modell mit guten vs. schlechten Entscheidungen zu trainieren. Konkret fällt das unter den „Framing Effect Bias".

Das Thema „Verzerrungen" wird vermutlich in Zukunft stärker Thema werden, auch bezogen auf die Erstellung von Trainingsdaten, die möglichst repräsentativ und eben „verzerrungsfrei" sein sollen.

Bei all den Verzerrungen kann man sich fragen, ob es überhaupt Sinn ergibt, KI einzusetzen. Menschen sind allerdings auch nicht frei von Verzerrungen. Sie verwenden viele Daumenregeln (Heuristiken) und entsprechend weist das Denken viele Verzerrungen auf. All diese Heuristiken und Verzerrungen sollten aber nicht darüber hinwegtäuschen, dass Menschen relativ gut darin sind, Vorhersagen und Entscheidungen zu treffen – es ist schlicht überlebenswichtig. Es ist aber auch nicht verwunderlich, dass ein KI-System – die von Menschen entwickelt wurde – auch sehr leicht Verzerrungen aufweisen kann.

Wie kann man Verzerrungen (biases) vermeiden? Srinivasan und Chander (2021) geben die Empfehlungen, dass domänenspezifisches Wissen zentral ist und die Eigenschaften, anhand derer die KI lernen soll, bewusst ausgewählt werden müssen. Außerdem muss die Datenbasis der Population entsprechen, es muss klare Standards geben, wie die Daten annotiert werden (labels), und die relevanten Variablen müssten klar identifiziert werden (u. a. auch welche fehlen). Schließlich muss das Modell mit einer repräsentativen Stichprobe getestet werden.

11.4.3 Datenschutz & Privatsphäre

KI-Systeme – insbesondere wenn maschinelles Lernen verwendet wird – benötigen Daten. Diese Daten müssen allerdings „irgendwoher kommen“. Und dabei müssen Datenschutz und Privatsphäre gewahrt werden, speziell (via Poretschkin et al., 2021) das Recht auf Informationelle Selbstbestimmung sowie die Persönlichkeitsrechte. Das ist insbesondere bei Stimmaufnahmen aber auch Fotos und Videos relevant. Rechtlich sind u. a. die Datenschutz-Grundverordnung (DSGVO) sowie das Bundesdatenschutzgesetz (BDSG) ausschlaggebend. Bei der Verwendung von Daten ist zu beachten, dass es nicht darum geht, dass ein Missbrauch tatsächlich passiert ist. Es ist schon problematisch, wenn dieser möglich wäre.

Datenschutz und Privatsphäre gehen allerdings über das reine Abgreifen von Daten durch unbefugte Personen hinaus („data leak“). Ein weiteres Problem ist, dass beim Trainieren von KI-Systemen oft Daten zusammenführt werden („data linkage“). Durch diese Zusammenführung von Daten können leichter Personen identifiziert werden, was – bei einem Datenleck – schnell zu Rufschädigung bzw. finanziellen Schäden führen kann.

Auf der anderen Seite ist ein überzogener Fokus auf Datenschutz und Privatsphäre eine Innovationsbremse. Das heißt nicht, dass man Datenschutz und Privatsphäre ignorieren sollte – oder dass es überbewertet wäre. Aber man muss hier auch einen Blick für die Konsequenzen haben. Ein KI-System braucht für gute Entscheidungen Daten zum Lernen – das ist bei einem Menschen nicht anders. Menschen entwickeln ihre Expertise im Umgang mit vielen Fällen.

Um Datenschutz und Privatsphäre zu schützen, ist die Einwilligung der Betroffenen notwendig, die Weiterverarbeitung darf nur mit Zustimmung erfolgen, es darf keine unberechtigten Zugriffsmöglichkeiten geben und es muss ein weitreichendes und jederzeitiges Widerspruchsrecht gewährleistet werden. Personen müssen auch Information über Zweck und Einsatz der personenbezogenen bzw. daraus abgeleiteten Daten erhalten. Generell gilt das Prinzip der Datensparsamkeit und das Prinzip der zweckgebundenen Verwendung. Diese Kriterien müssen berücksichtigt werden. Mehr Informationen dazu im Kap. 12 („KI & Recht“).

11.4.4 Vermeidung von (oft subtilen) Beeinflussungen

Der Einfluss von KI kann auch subtiler sein – das Verhalten verändert sich, weil es auf bestimmte Kriterien hin optimiert wird. Besonders sichtbar ist dies derzeit bei sozialen Medien (vgl. zum Beispiel Sherman et al., 2016).

Ein KI-System kann mit einer eigenen Agenda einhergehen – mit einem beständigen Druck, der ausgeübt wird (vgl. die „Stärken" von Computern, wenn Personen beeinflusst werden, in Abschn. 11.3, bezüglich „Persuasive Technology").

11.4.5 Auswirkungen von KI auf die öffentliche Verwaltung

Von den Beeinflussungen etwas weiter gedacht ist auch die Frage, wie sich die öffentliche Verwaltung langfristig weiterentwickelt, wenn KI-Systeme eingesetzt werden. Viele Auswirkungen werden erst langfristig sichtbar werden, zum Beispiel:

- Wie verändern sich die wahrgenommene Autonomie, Kompetenz und sozialen Beziehungen bei der Arbeit?
- Ist tatsächlich mehr Zeit für die Bürger verfügbar, wenn KI-Systeme zeitintensive, einfache Aufgaben übernehmen – oder entstehen neue Aufgaben?

Es liegen schon heute konkrete Befürchtungen vor, wie z. B. Stellenabbau, Fertigkeitsabbau („de-skilling", was insbesondere ein Problem wird, wenn die Automation versagt) oder der Erhalt der Aufmerksamkeit, wenn Handlungen seltener werden.

11.4.6 Gesellschaftliche Auswirkungen

Der Einfluss von KI-Systemen auf soziale Gefüge ist sehr schwer vorherzusehen. Wann immer Technologie eingesetzt wird, verändern sich auch soziale Beziehungen – und umgekehrt. Ein passendes, wenn auch nicht KI-bezogenes Beispiel, ist die Einführung eines Mängelmelders. Bürger sehen Probleme (z. B. wilder Müll) und können dies direkt den zuständigen Stellen mitteilen. Allerdings können sie auch in umfangreichen Maß Fehlverhalten anderer melden. Dies kann Auswirkungen auf eine Gemeinschaft haben. Technik kann in sozialen Systemen immer zu unerwarteten Wirkungen führen. Auswirkungen müssen daher auch langfristig beobachtet und mit Blick auf die ursprünglichen Ziele bewertet werden.

11.4.7 Zielkonflikte

Bei der ethischen Bewertung sind mögliche Zielkonflikte zu berücksichtigen (vgl. auch die unterschiedlichen Definitionen von Fairness, Abschn. 11.4.1, oder die besonderen Anforderungen bei KI-Anwendungen, Abschn. 6.5).

Das ist nach Poretschkin und anderen (2021) z. B. der Fall bei Datensparsamkeit und Fairness. Um eine höhere Genauigkeit zu erzielen (für faire Entscheidungen), müssen in umfassenden Maß Daten erhoben werden (widerspricht dem Ziel der Datensparsamkeit). Die Genauigkeit – via eines gut trainierten Black-Box-Systems – kann auch leicht in Konflikt mit der Interpretierbarkeit kommen. Menschliche Aufsicht und Autonomie kann in Konflikt mit Sicherheit kommen, da Manipulation möglich ist. Letztlich sind Rest-Risiken und Trade-Offs oft unvermeidlich.

11.5 Ethische Bewertung von KI-Anwendungen

Wie kann eine KI-Anwendung ethisch bewertet werden – inwieweit werden ethische Standards eingehalten oder verletzt?

Als Einzelperson können in der Praxis oft nur Teilaspekte bewerten werden. Als Nutzer wird man auf frühe Aspekte der KI-Pipeline nur wenig Einfluss oder wenig Wissen dazu haben. Aber man kann sich z. B. ansehen, wie Entscheidungen getroffen werden und ob die Ergebnisse nachvollziehbar sind oder nicht. Auch langfristige Auswirkungen sind zumindest abschätzbar.

Insbesondere sollte man sich bei der Bewertung von KI-Anwendungen der Eigenverantwortung bewusst sein. Man kann sie weder an das KI-System noch an die Programmierer abgeben (vgl. Abschn. 11.3.7).

Es ist allerdings sehr leicht, im Umgang mit KI einfach nur Anweisungen zu befolgen, gerade weil das KI-System den Eindruck erwecken kann, als würden die Entscheidungen „objektiv" und „richtig" erfolgen. Wie in diesem Kapitel beschrieben, ist das nicht unbedingt der Fall.

In diesem Zusammenhang ist ein Exkurs zu einem fragwürdigen Experiment aus der Psychologie zu Gehorsam hilfreich: dem Milgram Experiment (vgl. Baron et al., 2006; Milgram, 1963).

Milgram wollte nach dem zweiten Weltkrieg wissen, warum so viele Deutsche einfach „nur Befehle befolgt" hatten. War es eine typisch deutsche autoritäre Persönlichkeit – wie von einigen Personen angenommen oder steckte mehr dahinter, etwas, was auch in anderen Ländern passieren könnte?

Milgram hat die Frage in einem Experiment untersucht, das kurz zusammengefasst wie folgt ablief. Er hat erwachsene Teilnehmer für seine Studie rekrutiert (z. B. Flyer), denen angeblich zufällig entweder eine Lehrer- oder Lerner-Rolle zugelost wurden. In Wirklichkeit war aber die Auslosung so manipuliert, dass der Teilnehmer immer der Lehrer war, während der Lerner in Wirklichkeit ein Helfer von Milgram war. Der Lerner wurde in einem Raum auf einen Stuhl festgebunden und ihm wurde etwas umgelegt, um ihm Elektroschocks zu geben. Der Lehrer (also der echte Versuchsteilnehmer) wurde in einen anderen Raum geführt und musste dem Lerner über eine Sprechanlage Wortpaare beibringen. Zuerst wurden die Wortpaare genannt, im weiteren Verlauf wurde dann nur das erste Wort gesagt und vier mögliche Lösungen vorgegeben. Eine davon war richtig, die anderen drei falsch. Wenn der Lerner eine falsche Antwort gab, sollte der Lehrer ihn mit einem Elektroschock bestrafen. Dafür gab es eine Maschine, die (angeblich) Elektroschocks von 15 bis 450 V geben konnte, jeweils in 15-V-Schritten. Macht der Lerner einen Fehler, sollte der Lehrer einen Schock geben – zuerst 15, dann 30, dann 45, etc. bis 450 V. In Wirklichkeit wurden keine Schocks gegeben, aber für den Lehrer sah es echt aus, inkl. der Schmerzreaktionen des Lerners über die Sprechanlage. Die sehr emotionalen Reaktionen des Lehrers haben auch gezeigt, dass die Lehrer dies geglaubt haben. Es gab mehrere Varianten des Experiments. In einer schlug der Lerner bei 300 V gegen die Wand, bat darum aufzuhören, beklagte Herzprobleme und antwortete dann nicht mehr. Der Versuchsleiter, der die ganze Zeit anwesend war, gab die Anweisung, dass keine Antwort als Fehler zu werten ist. Wenn der Lehrer unsicher war, ob er weitermachen sollte oder sich weigerte, gab der Versuchsleiter die Anweisung weiter zu machen, z. B. mit „Bitte machen Sie weiter.“, „Das Experiment erfordert, dass Sie weiter machen.“, „Es ist absolut notwendig, dass Sie weiter machen.“, oder „Sie haben keine andere Wahl, Sie müssen weiter machen.“.

Die Frage war, befolgen die Lehrer – normale Menschen die an der Studie teilnahmen – die Anweisungen und geben potenziell tödliche Elektroschocks? Milgram hatte vorher Psychiater befragt. Deren Einschätzung war, dass die Lehrer früh aufhören würden. Fast keiner würde über 300 V gehen und nur in Ausnahmefällen würden die Versuchspersonen bis zum Ende gehen.

In dem beschriebenen Experiment sind 26 von 40 Teilnehmern (65 %) bis zum Ende gegangen, also bis zu 450 V – eine Schockstärke, die potenziell tödlich ist, und nachdem der Lerner ab 300 V nicht mehr reagiert hat. *Keiner* hat vor 300 V aufgehört.

Auch wenn das Experiment umstritten ist, hat es das Weltbild von einem Charakterfehler einer Minderheit von Personen die „nur Befehle befolgen“ infrage gestellt.

Milgram hat untersucht, was passiert, wenn der Lehrer im selben Raum wie der Lerner ist (es hören mehr Personen früher auf) oder wenn ein zweiter Lehrer mit im Raum ist (in Wirklichkeit ein weiterer Mitarbeiter von Milgram), der sich widersetzt (es hören mehr Personen früher auf). Milgram hat auf diese Weise Faktoren identifiziert, die Gehorsam bei unethischen Anweisungen wahrscheinlicher oder unwahrscheinlicher machen (vgl. Baron et al., 2006; Milgram, 1963).

Was geschieht, wenn diese Faktoren auf den Einsatz von KI-Systemen übertragen werden? In Klammern werden die Faktoren aus dem Experiment genannt. Ist es denkbar, dass Systemanwender unethische Entscheidungen akzeptieren, wenn das System klar entscheidet (der Versuchsleiter weist an), dem System ein objektiver und fairer Status durch ein Zertifikat zugeschrieben wird (Versuchsleiter als Autoritätsperson durch Insignien, z. B. Laborkittel), das System schrittweise verstärkt unethisch entscheidet (schrittweise Erhöhung der Stromschläge) und das System unter Zeitdruck genutzt wird? Ein Faktor der Ungehorsam wahrscheinlicher macht, also eine Ablehnung von unethischen Entscheidungen bedeutet, kann zum Beispiel das Hinterfragen des Systems sein (Hinterfragen der Expertise, Motive und des Wissens der Versuchsleitung).

Im unreflektierten Einsatz von KI besteht die Gefahr, dass Verantwortung dem KI-System zugeschrieben wird, dass die KI-Anwendung als fehlerfrei wahrgenommen wird, dass der Einsatzbereich des KI-Systems schleichend erweitert wird, Nutzer von der Entwicklung überrollt werden (z. B. wegen hoher Arbeitsbelastung) und letztlich niemand widerspricht.

Vor diesem Hintergrund müssen KI-Systeme anhand klarer Kriterien bewertet werden (vgl. Abschn. 11.7).

11.6 Ethische Aspekte von KI – Interview mit Christian Herzog, Leiter des Ethical Innovation Hubs der Universität zu Lübeck

Eine interessante Perspektive auf ethische Aspekte von KI bietet ein Interview, das Ende 2021 mit Christian Herzog, dem Leiter des Ethical Innovation Hubs der Universität zu Lübeck durchgeführt wurde. Er verbindet die ingenieurwissenschaftliche/informatische Perspektive mit einer ethischen Perspektive, um ethische Analysen nutzbringend für die technische Entwicklung einzusetzen. Im Folgenden sind seine Antworten auf die Interviewfragen zusammengefasst und leicht gekürzt wiedergegeben.

Wie würdest du kurz und prägnant definieren, was Ethik ist?
Ethik ist die wissenschaftliche Beschäftigung mit unserem Moralverständnis, die deskriptiv (Was ist? Wie verstehen wir? Wie handeln wir? Was denken wir ist gut und richtig?) oder normativ (Was sind Prinzipien, die helfen herauszufinden, was gut und richtig ist? Wie können wir unser Moralverständnis hinterfragen?) erfolgen kann. Sie ist besonders bei KI wichtig, weil wir uns mit neuen Entwicklungen konfrontiert sehen. Wir greifen zwar auf bisherige Entwicklungen zurück und nutzen unser Moralverständnis, aber wir müssen auch in die Zukunft schauen. Ein weiterer Bereich ist die angewandte Ethik, bei der bei ganz konkreten Fragestellungen (Was soll ich der Situation machen?) Abwägungen und Argumente bereitgestellt werden, um die Entscheidung treffen zu können.

Aus welcher Perspektive betreibst du Ethik? Welche Aspekte von Ethik sind dir besonders wichtig?
Insbesondere die angewandte Ethik. In der Medizintechnik bzw. medizinischer Informatik stellt sich z. B. die Frage, inwieweit KI erklärbar sein muss. Insbesondere wenn Systeme schon sehr gut funktionieren, möchte man diese den Patienten nicht vorenthalten. Auf der anderen Seite sind langfristige Perspektiven, was z. B. den Erkenntnisgewinn in der Medizin betrifft oder die Auswirkungen auf Patient-Arzt Beziehung zu beachten.

In der öffentlichen Verwaltung ist es u. a. die Frage, mit welchen Prioritäten KI-Methoden eingesetzt werden. Wer verwendet die KI, wer ist davon betroffen (profitiert, wird davon beeinflusst) und aus welchen (eher grundlagenphilosophischen) Beweggründen wird sie eingesetzt. Warum digitalisieren wir die öffentliche Verwaltung und setzen KI ein? Da ist die Sichtweise von Jürgen Habermas interessant, der vielleicht idealisiert oder überhöht formuliert eine ideale Sprechsituation beschreibt. Sie ist auf Augenhöhe, ohne Machtgefälle, mit vollkommender Offenheit bei der das beste Argument gewinnt. Beim Einsatz der KI benötigen wir eine breit angelegte Partizipation. Und die benötigt Zeit. Wir müssen uns Mühe geben und die Zeit investieren mit allen Betroffenen, diese gemeinsam an einen Tisch zu bringen. Vielleicht in Stufen, aber wir brauchen eine demutsvolle Perspektive um Konsens zu bewirken, den politischen Prozess anzustoßen. Statt im Elfenbeinturm auszuarbeiten wie man vorgehen muss sollte man einen Aushandlungsprozess anstoßen.

Welche ethischen Aspekte sind bei KI gesellschaftlich besonders relevant?
Bei der gesellschaftlichen Perspektive sind wir nicht mehr ganz nah an der Technologie oder am konkreten Gegenstand. In Prinzipienkatalogen, die man kritisch sehen kann, wird zuerst immer der Respekt vor der menschlichen Autonomie genannt.

Wenn wir die respektieren, dann müssen wir Gestaltungsspielräume geben oder erhalten. Diese werden immer eingeengt durch Gestaltungsspielräume anderer Personen, aber bei KI muss man (auch langfristig) sehen, ob der Gestaltungsspielraum eingeengt wird. Derzeit haben wir durch KI einen großen Nutzen, aber auf lange Sicht machen wir uns abhängig von einer solchen Technologie. Wir haben auch eine solche Durchdringung dieser Technologie, dass wir sie kaum noch zurücknehmen können. Dies sehen wir in vielerlei Hinsicht bei Smartphones. Wir können wahrscheinlich ganz gut damit leben, wenn wir privilegiert genug sind uns mit Zeit und Muße damit auseinander zu setzen und uns wenn notwendig von der Technik distanzieren. Dann sind wir davon so abhängig wie von Papier und Bleistift. Aber das muss nicht bei jedem so sein. Entsprechend ist der wichtigste und schwerwiegendste Punkt eine langfristige Perspektive einzunehmen und es auch in Partizipationsformaten zu schaffen, dass Menschen diese langfristige Perspektive gelten lassen.

Im Gegensatz zur Medizin, wo es direkt um Menschenleben geht und Erklärbarkeit auf praktischen Nutzen trifft, haben wir diesen Druck in der öffentlichen Verwaltung häufig nicht. Die Prozesse laufen, vielleicht nicht ideal, aber wir können uns Zeit und Muße nehmen in einen partizipativen Vorgang einzutreten. Wir können Mitarbeiter der Verwaltung mitnehmen, langfristig denken, auch z. B. Umschulungsmaßnahmen mit einzubeziehen, um den (auftretenden) Wegfall von Arbeitsplätzen sozial gerecht zu gestalten. Das sind Fragen, die wir gar nicht so direkt am technischen Artefakt beantworten müssen, sondern im gesellschaftlichen Kontext.

Welche ethischen Aspekte sind bei KI in der öffentlichen Verwaltung besonders relevant?

Wir dürfen Ermessensspielraum nicht „aus Versehen“ verringern, was schleichend passieren kann. Wir müssen diese Entwicklungen ernst nehmen, auch aus psychologischer Sicht. Zum Beispiel werden Entscheidungsunterstützungssysteme aus guter Intention heraus entwickelt. Man versucht, objektivere, bessere, mit Informationen gestützte Entscheidungen herbeizuführen. Es gibt Effekte, wenn Systeme nicht ganz durchschaut und täglich verwendet werden, dass man ihnen übermäßig vertraut. Empfehlungen werden als de facto Entscheidungen umgesetzt. Damit wird der Entscheidungsspielraum auf null reduziert. Das kann nicht das Ziel von Digitalisierung oder KI sein. Es wird gesagt, dass durch KI mehr Interaktion mit Bürgern stattfinden soll (politische Frage). Wenn Technologie den Freiraum schaffen kann, dann darf man ihn nicht auf der anderen Seite durch Rationalisierung wieder einengen.

Die Unterstützung, welche die KI leisten kann, kann korrumpiert werden von Dingen, für die die Technologie erstmal nichts kann. Deswegen muss man global und umfassend in der öffentlichen Verwaltung über KI nachdenken. Es ist weniger

eine Frage der KI selbst, sondern wie sie eingebettet ist – das Interface Design, oder wie die KI in den Arbeitsfluss in der Verwaltung eingebettet ist. Dass man z. B. darauf hinweist, dass es nur eine Empfehlung ist und wo man nachschlagen muss, was die Entscheidungskriterien betrifft. Auch den Aspekt, was versprochen wurde und was umgesetzt wurde, muss man beachten. Wir bewegen uns im Kontext was in den 80er Jahren über Automatisierung gesagt wurde.

Die Ironie der Automatisierung war, dass die Arbeit durch die Arbeitsverdichtung unmenschlich wurde, wenn am laufenden Band acht Stunden am Stück gearbeitet wurde. Wenn durch den Einsatz von KI nur noch die schweren Fälle übrig bleiben, stellt sich die Frage, wie dann der Feierabend noch aussehen kann und wie emotional belastet man dann ist. Da muss man sich die Arbeitsprozesse ansehen. Wir sagen zwar, wir wollen nicht monoton arbeiten, aber wenn monotone Phasen nicht mehr vorhanden sind, würden wir dann mehr arbeiten? Vermutlich würden wir mehr Kaffee trinken.

Da ist das gute Bild vom Menschen, der auch leistungswillig ist und Teil der Gesellschaft ist. Der Teil eines funktionierenden Apparates ist, und nicht gegen den Bürger, der hart verwaltet werden muss, sondern bei dem ein Miteinander im Vordergrund steht.

Wie wird KI die öffentliche Verwaltung verändern? Worauf muss die öffentliche Verwaltung bei Ethik besonders achten, damit es nicht zur Dystopie wird?

Vielleicht zunächst wie wir an einer Dystopie vorbeikommen. Auch auf die Gefahr hin, jetzt erzkonservativ zu wirken, aber wir müssen uns ansehen, was als nächster Schritt angedacht ist. Robotic Process Automation (RPA, vgl. Kap. 9) eignet sich sehr gut. Das wird auch in Gesetzesvorlagen einbezogen, da kein Ermessensspielraum beeinflusst wird. Es kann zwar aus Versehen passieren, aber mit RPA sind wir auf einem sicheren Weg, dass dies erst mal nicht passiert und wir Erfahrungen sammeln können. Effizienzsteigerungen, die man erhält und erhalten darf, erst einmal zu verwalten. Sich ansehen, was mit der Arbeitskultur passiert, die sich auf das System einstellen muss. Und den Prozess mit zu begleiten, die Mitbestimmung der Angestellten sicherzustellen, zusehen, dass die Transition gut verläuft und das Bürger davon profitieren. Es ist weniger Dystopie, sondern was ich mir wünschen würde ist, dass digitale Souveränität zum moralischen Kompass für Digitalisierung in der Verwaltung wird. Dass die Souveränität von Bürgern gesteigert wird, nicht in einem Kunden-Serviceverhältnis aus Bürger und Verwaltung, sondern durch den alltäglichen Kontakt mit der Verwaltung. Wir müssen informieren, Partizipation auf niedrigschwellige Art sicherstellen, Vertrauenswürdigkeit und Akzeptanzfähigkeit herstellen. Die Möglichkeit schaffen, dass es akzeptiert wird, begründet darauf, dass

man weiß, was das Gegenüber für Ziele hat und auch meine Ziele – also hier die der Bürger – respektiert.

Als Dystopie – eine Idee aus der Wissenschaft, die sich mit Privatheit beschäftigt, kommt von Daniel Solove. Er beschäftigt sich mit Digitalisierung, Datenaustausch zwischen Behörden, dem Ansammeln von Daten. Privatsphäre wird manchmal unter der Drohgebärde von „Big Brother" thematisiert, was Solove für falsch hält. Er sieht das Problem eher wie bei „Der Prozess" von Franz Kafka. Es geht darum, dass viele sehr bürokratische, unmenschliche einzelne Entitäten in der Verwaltung Informationen austauschen, ohne Wissen des Betroffenen und auch ohne Intention zu überwachen oder Repressalien anzuwenden. Aber es funktioniert einfach ineinander in einer uneinsehbaren Art, es werden Daten interpretiert und man wird nur noch mit Ergebnissen konfrontiert. Das wäre eine wesentliche Dystopie.

Man sollte nicht den Fehler machen, aus der kommerzialisierten Welt Analogien zu ziehen, aber dort ist auch nicht alles falsch. Eine Bank stellt z. B. einen persönlichen Berater zur Verfügung, mit dem man immer wieder spricht. Wenn KI die Ressourcen frei macht, dann kann sich das die öffentliche Verwaltung vielleicht auch leisten. Ein persönlicher Draht, der den Austausch effizienter macht, weil man die Lebensgeschichte nicht immer neu erzählen muss. Es wäre mehr Empathie möglich, man könnte den Ermessensspielraum nutzen ohne ungerecht zu werden oder bevorteilen zu wollen. Man kann entscheiden, ob man sich mehr Mühe geben muss oder ob jemand alleine zurechtkommt.

Was würdest du Mitarbeitern der ÖV mitgeben, bzw. Personen, die mit KI zu tun haben, bezüglich Ethik und KI?

Wenn man auf Partizipation abzielt, dann verlangt es einiges von Mitarbeitern der öffentlichen Verwaltung, eine gewisse Offenheit, Neugier, und Spaß, dass sich was verändert. Mir ist bewusst, dass man den Spaß nicht in jeder Lebenssituation haben kann. Man kann sich aber organisieren und Partizipation einfordern. Es gibt Stellen, die Lobbyarbeit machen, dass die Digitalisierung der öffentlichen Verwaltung nicht an den Mitarbeitern vorbeigeht. Sie muss auf fruchtbaren Boden fallen. Ich würde es mir wünschen, weil ich glaube, dass man dann den Bereich, der manchmal zu Unrecht bei Bürgern negative Konnotation aufweist („die Verwaltung", „muss wieder auf's Amt") einen anderen Anstrich geben würde. Einen, der gerechtfertigt ist, weil dort in Zukunft moderne Prozesse ablaufen, die hoffentlich viel Geld sparen. Das kommt den Bürgern zu Gute. Ich würde mir wünschen, dass Mitarbeiter sich einsetzen und aufbegehren, dass der Arbeitsplatz weiterhin einer ist, den sie bekleiden wollen. Man muss in der Gesellschaft zusammenhalten und nicht sagen, dass es die Sache der Mitarbeiter der Verwaltung ist. Bürger müssen das auch im eigenen Interesse einfordern – die Verwaltung profitiert, aber auch die Bürger. Es

geht nicht so sehr darum, was die KI ausmacht – das ist technisch faszinierend und es ist auch schön, wenn man sich dafür interessiert. Es geht eher darum, was die Ziele der Leute sind, die Digitalisierung maßgeblich anstoßen, dass man kritisch bleibt, nachfragt, mitgestaltet und eigene Ziele einbringt und in den Diskurs kommt. Dann bin ich zuversichtlich, dass es zu einem guten Einsatz kommt.

Das Problem in der KI-Entwicklung ist, dass in der Regel von Menschen entwickelt wird, die nicht davon betroffen sind. Nicht aus bösem Willen, sondern weil es schwer ist, an Menschen heranzukommen, die davon betroffen sind. Die leben z. B. in einer prekären Situation, müssen arbeiten, oder lehnen Technologie aus einer Grundhaltung ab, die nicht unbegründet sein muss. Man muss auch Partizipation ernst nehmen – Anreizstrukturen schaffen, Erleichterungen für Leute schaffen, die üblicherweise am Prozess nicht teilnehmen. Die Gründe muss man erst nehmen und Angebote machen für eine diverse Perspektive auf den Wandel, der nicht nur die technologie-affinen und Zukunftsutopisten anspricht. Es ist schön diese zu haben, man braucht aber auch kritische Stimmen.

Als letzte Frage, gibt es noch weitere Punkte, bezüglich Ethik und KI, die wir nicht angesprochen haben, die aber wichtig sind?
Ein Thema, das heißt diskutiert wird und wir nicht angesprochen haben, ist „algorithmic bias“. Entscheidungsunterstützungssysteme, die auf verzerrten Daten basieren oder von Entwicklungsteams mit bestimmter Perspektive entwickelt wurde – und die eine andere Perspektive auch bei besten Intentionen nicht einnehmen können. Da wird dem KI-System eine bestimmte Perspektive mitgegeben auch ohne es zu wollen, oder auch aus Nachlässigkeit. Gerade bei Fragen in der öffentlichen Verwaltung kann das massiv soziale Ungerechtigkeit hervorrufen. Auch an scheinbar banalen Entscheidungen können Schicksale dranhängen, wenn es ums Geld geht. Das Thema kann man nicht unterbewerten, das ist sehr wichtig und das müssen wir bei KI berücksichtigen.

11.7 Fragen an KI-Anwendungen in der öffentlichen Verwaltung

Was sind Fragen, die man sich bei KI-Anwendungen in der öffentlichen Verwaltung stellen kann? Wie kann man die ethischen Aspekte von KI-Anwendungen überprüfen? Siehe dazu auch Unterkapitel 6.7 und 7.7.

Fairness

- Ist die Anwendung fair – nach welcher Definition von Fairness?
- Werden unverzerrte, faire, Entscheidungen getroffen?

Langfristige Auswirkungen durch den Einsatz von KI

Eigenes Arbeitsverhalten

- Beeinflussung: Verändert sich das individuelle Arbeitsverhalten über die Zeit negativ? Wird es durch den Einsatz von KI in eine bestimmte Richtung gedrückt?

Auswirkungen auf die öffentliche Verwaltung

- Wie verändern sich die wahrgenommene Autonomie, Kompetenz und sozialen Beziehungen bei der Arbeit?
- Wurden die Versprechungen/Hoffnungen des KI-Einsatzes auch eingelöst? Gab es negative Konsequenzen (z. B. Stellenabbau, Fertigkeitsabbau)?

Auswirkungen auf die Gesellschaft

- Welche Auswirkungen hat der Einsatz von KI in der öffentlichen Verwaltung auf die Gesellschaft als Ganzes?

Einsatz der KI-Anwendung (Organisationsebene, Wing, 2021; Shneiderman, 2021)

- **Datenbasis:** Wurden die Daten auf ethische Weise gesammelt?
- **Nachhaltigkeit:** Wie ist die Kosten-Nutzen-Rechnung bezüglich Energieverbrauch/Nachhaltigkeit?
- **Verzerrungsfreie ML-Pipeline:** Wurden Verzerrungen in der ML-Pipeline ausgeschlossen/reduziert?
- **Verlässlichkeit:** Wird der Einsatz der KI-Anwendung regelmäßig überprüft? Z. B. via:
  - Dokumentation des Verhaltens der KI-Anwendung
  - Audits, Benchmark Tests
  - Verwendung von Analyse-Werkzeugen, Erklärbare KI-Ansichten
  - Kontinuierliche Begutachtung der Datenqualität und Testen auf mögliche Verzerrungen

- **Sicherheit:** Ist die Sicherheit auf organisatorischer Ebene gewährleistet? Z. B. via:
  - Verpflichtung zur Sicherheit durch Führungskräfte
  - Offenes Berichten über Fehler und kritische Ereignisse
  - Öffentliche Berichte von Problemen und zukünftigen Plänen
- **Vertrauenswürdigkeit:** Wird die Vertrauenswürdigkeit auf organisatorischer Ebene gewährleistet? Z. B. via:
  - Einhalten von Standards und Richtlinien
  - Zertifizierung
  - Externe Kontrolle

11.8 Übungsfragen: KI & Ethik

Zur Überprüfung Ihres Wissensstandes können Sie die folgenden Fragen beantworten.

1. Was ist Ethik?
2. Warum ist Ethik bei KI besonders relevant?
3. Welche Hauptaspekte gibt es bei KI & Ethik?
4. Wann ist eine Entscheidung fair?
5. Welche Verzerrungen können über die ML-Pipeline auftreten?
6. Was muss bei Datenschutz und Privatsphäre sichergestellt werden?

11.9 Aufgaben zum eigenen Anwendungsfall

Sie haben die Anwendung bereits auf Gebrauchstauglichkeit und im Detail bezüglich der Mensch-KI-Interaktion analysiert. Schauen Sie jetzt auf die ethischen Aspekte von KI-Anwendungen.

- Inwiefern ist die Anwendung fair – insbesondere für die von den Entscheidungen betroffenen Personen? Welche Definition von Fairness verwenden Sie? Woran könnten Sie etwaige Probleme erkennen? Zusätzlich, falls maschinelles Lernen verwendet wird: Wie stellen Sie sicher, dass Verzerrungen in der ML-Pipeline ausgeschlossen bzw. reduziert wurden?

- Welche langfristigen Auswirkungen durch den Einsatz der KI-Anwendung sind realistisch? Nehmen Sie dann eine kritische Perspektive ein – was könnte schlimmstenfalls passieren (Beeinflussung, Auswirkungen auf die öffentliche Verwaltung, Auswirkungen auf die Gesellschaft) und wie sollte sichergestellt werden, dass diese Probleme nicht eintreten?
- Gehen Sie im letzten Schritt über Ihre KI-Anwendung hinaus und beurteilen Sie den Einsatz von KI-Anwendungen in Ihrer Verwaltungseinheit generell. Wie wird auf Organisationsebene (Einsatz der KI-Anwendung auf Organisationsebene) vorgegangen? Ist das Vorgehen ethisch?

Literatur

Baron, R. A., Byrne, D., & Branscombe, N. R. (2006). *Social Psychology* (11. Aufl.). Pearson Education, Inc.

Fogg, B. J. (2003). *Persuasive technology: Using computers to change what we think and do.* Morgan Kaufmann.

IBM. (2022). *Design for AI*. IBM Corp. https://www.ibm.com/design/ai/ethics/value-alignment. Zugegriffen: 10. Okt. 2022.

Milgram, S. (1963). Behavioral study of obedience. *The Journal of Abnormal and Social Psychology, 67*(4), 371–378. https://doi.org/10.1037/h0040525.

Poretschkin, M., Schmitz, A., Akila, M., Adilova, L., Becker, D., Cremers, A., Hecker, D., Houben, S., Mock, M., Rosenzweig, J., Sicking, J., Schulz, E., Voss, A., & Wrobel, S. (2021). *Leitfaden zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz–KI-Prüfkatalog. Fraunhofer IAIS.* https://www.iais.fraunhofer.de/ki-pruefkatalog. Zugegriffen: 14. Juli 2021.

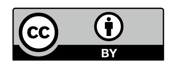
Sherman, L. E., Payton, A. A., Hernandez, L. M., Greenfield, P. M., & Dapretto, M. (2016). The power of the like in adolescence: Effects of peer influence on neural and behavioral responses to social media. *Psychological Science, 27*(7), 1027–1035.

Shneiderman, B. (2021). Responsible AI: Bridging from ethics to practice. *Communications of the ACM, 64*(8), 32–35. https://doi.org/10.1145/3445973.

Srinivasan, R., & Chander, A. (2021). Biases in AI systems. *Communications of the ACM, 64*(8), 44–49. https://doi.org/10.1145/3464903.

Verma, S., & Rubin, J. (2018). Fairness definitions explained. *Proceedings of the International Workshop on Software Fairness,* 1–7. https://doi.org/10.1145/3194770.3194776.

Wing, J. M. (2021). Trustworthy AI. *Communications of the ACM, 64*(10), 64–71. https://doi.org/10.1145/3448248.


BY



12 KI & Recht

Zusammenfassung

In diesem Kapitel werden die rechtlichen Rahmenbedingungen rund um den Einsatz von KI-Systemen aufgeführt. Zunächst werden erste Regulierungsansätze vorgestellt (12.2). Anschließend werden die grundsätzlichen Prinzipien der Datenschutz-Grundverordnung beschrieben (12.3). Mittels einer kleinen Übung wird das Bewusstsein zur Kategorisierung personenbezogener Daten geschärft (12.4). Es folgt eine Diskussion über die Vereinbarkeit von KI-Systemen und Datenschutz am Beispiel des Einsatzes von Chatbots (12.5). Schließlich wird der rechtlich anspruchsvolle Anwendungsfall des vollständig automatisierten Verwaltungsakts analysiert (12.7). Das angeeignete Wissen kann in der Übung reflektiert (12.7) und im eigenen Anwendungsfall eingesetzt werden (12.8).

12.1 Einleitung

KI-basierte Systeme werden in absehbarer Zukunft auch im öffentlichen Sektor eine immer größere Rolle spielen. Allerdings muss sich insbesondere die öffentliche Verwaltung beim Einsatz neuer Technologien immer eine Frage stellen: Dürfen wir das auch? Technische Innovationen haben immer einen begrenzenden Faktor: das Gesetz. Denn jedes Verwaltungshandeln basiert auf Recht und Gesetz oder ist zumindest gesetzlich umrahmt. Weder das Grundgesetz noch das Verwaltungsverfahrensgesetz macht der Verwaltung zwar eine generelle Vorgabe hinsichtlich der zur Aufgabenerfüllung zu nutzenden Instrumente. Doch unsere Rechtsordnung ist anthropozentrisch aufgebaut, das heißt der Mensch steht im

M. Heine et al., *Künstliche Intelligenz in öffentlichen Verwaltungen*, Edition eGov-Campus, https://doi.org/10.1007/978-3-658-40101-6_12

Mittelpunkt des Regelungsansatzes, und auch die Anwendung des Rechts ist Aufgabe des Menschen (Hähnchen et al., 2020). Zudem gilt für alle Formen des Verwaltungshandelns die Gesetzesbindung an Art. 20 Abs. 3 GG:

▶Die Gesetzgebung ist an die verfassungsmäßige Ordnung, die vollziehende Gewalt und die Rechtsprechung sind an Gesetz und Recht gebunden.

In diesem Kapitel geht es um Regulierungsansätze für Künstliche Intelligenz in Europa und Deutschland. Außerdem wird analysiert, wie sich die Beziehung zwischen Datenschutz und dem Einsatz von KI-Systemen gestaltet und was unter personenbezogenen Daten zu verstehen ist. Schließlich wird noch vertieft das Thema des vollständig automatisierten Verwaltungsakts erörtert.

12.2 Erste Regulierungsansätze

Bislang gibt es auf nationaler Ebene nahezu keine Regelung, die explizit auf KI-Systeme ausgerichtet ist. Wie zuvor geschildert, greift die DSGVO, sobald personenbezogene Daten verwendet werden – ein darüberhinausgehender Regulierungsrahmen existiert hingegen nicht. Gleiches gilt, wenn man die europäische Ebene anschaut. Auch in der Europäischen Union wurde bislang noch keine Verordnung für KI-Systeme implementiert. Der Einsatz von KI wird daher aktuell lediglich durch einige nationale Gesetze und die DSGVO reguliert. Im April 2021 hat die EU-Kommission allerdings einen Entwurf für eine entsprechende Verordnung veröffentlicht, dieser wird seither intensiv diskutiert. Grundlage für den Entwurf war ein zuvor von der EU-Kommission publiziertes Weißbuch (Europäische Kommission, 2020).

Der Verordnungsentwurf der EU-Kommission vom 21. April 2021

Der Entwurf ist als ein Rechtsrahmen für KI anzusehen, der sich insbesondere mit den Risiken von KI befasst. Zweck der Verordnung ist es, durch die Etablierung eines einheitlichen Rechtsrahmens die Entwicklung, Vermarktung und Nutzung von KI im gemeinsamen Binnenmarkt zu verbessern. Außerdem soll der Schutz der Gesundheit, Sicherheit und Grundrechte im Kontext von grenzüberschreitendem Verkehr von KI-basierten Waren und Dienstleistungen sichergestellt werden. In Artikel 5 beschreibt der Entwurf Praktiken, die aufgrund eines nicht zu akzeptierenden Risikos verboten sind. Dazu gehören unter anderem:

- Systeme, die mithilfe unterschwelliger Techniken eine Person wesentlich beeinflussen oder physischen oder psychischen Schaden zufügen können;

- Systeme zur Bewertung der Vertrauenswürdigkeit eines Menschen auf Grundlage des Sozialverhaltens, sofern diese Systeme von Behörden genutzt oder in deren Auftrag in Betrieb genommen wurden;
- Echtzeit-Fernerkennungssysteme, die zur biometrischen Identifizierung einer Person im öffentlichen Raum zwecks Strafverfolgung verwendet werden.

Im Hinblick auf Echtzeit-Fernerkennungssysteme sieht der Entwurf aber Ausnahmen vor, in denen der Einsatz eines solchen Systems doch möglich ist. Eine solche Ausnahme ist etwa die gezielte Suche nach bestimmten potenziellen Opfern von Straftaten oder nach vermissten Kindern. Wenn eine konkrete, erhebliche und unmittelbare Gefahr für das Leben oder die körperliche Unversehrtheit einer natürlichen Person besteht oder ein Terroranschlag vorliegt, kann es ebenfalls zu Ausnahmen kommen. Ferner können KI-Systeme genutzt werden, um Täter oder Verdächtige im Sinne des Artikel 2 Absatz 2 des Rahmenbeschlusses 2002/584/JI des Rates zu erkennen, aufzuspüren oder zu verfolgen. Nach dem Beschluss handelt es sich hierbei um Personen, gegen die in dem betreffenden Mitgliedstaat nach dessen Recht eine Freiheitsstrafe oder eine freiheitsentziehende Maßregel der Sicherung von mindestens drei Jahren angesetzt ist.

Der Entwurf der EU-Kommission unterscheidet KI-Systeme, mit deren Nutzung ein hohes Risiko einhergeht von denen, die ein niedriges Risiko darstellen. Es wird von einem hohen Risiko ausgegangen, wenn ein KI-System in den folgenden Bereichen eingesetzt wird:

- Kritische Infrastrukturen (z. B. der Energiesektor oder das Gesundheitswesen);
- Schul- oder Berufsausbildung, wenn der Zugang einer Person zur Bildung oder zum Berufsleben dadurch beeinträchtigt werden kann (z. B. Bewertung von Prüfungen);
- Sicherheitskomponenten von Produkten (z. B. eine Roboterassistenz im Bereich der Chirurgie);
- Beschäftigung, Personalmanagement und Zugang zu selbstständiger Tätigkeit (z. B. Software zur Auswertung von Lebensläufen während eines Einstellungsverfahrens);
- Bestimmte private und öffentliche Dienstleistungen (z. B. die Bewertung der Kreditwürdigkeit);
- Strafverfolgung (z. B. Bewertung der Verlässlichkeit von Beweismitteln);
- Migration, Asyl und Grenzkontrollen (z. B. Überprüfung der Echtheit von Reisedokumenten);

- Rechtspflege und demokratische Prozesse (z. B. Systeme, die Justizbehörden dabei helfen sollen, Sachverhalte und Rechtsvorschriften zu ermitteln und auszulegen).

Dem letzten Punkt ist hinzuzufügen, dass ein KI-System, welches für rein begleitende Verwaltungstätigkeiten eingesetzt wird und die tatsächliche Rechtspflege nicht beeinträchtigt – wie etwa die Anonymisierung gerichtlicher Urteile –, nicht mit einem hohen Risiko assoziiert wird. KI-Systeme, deren Einsatz mit einem hohen Risiko verbunden wird, sollen vor der Marktzulassung strenge Anforderungen erfüllen:

- angemessene Risikobewertungs- und Risikominderungssysteme (Art. 9);
- hohe Qualität derjenigen Daten, mit denen das KI-System betrieben wird, insbesondere um Diskriminierungen zu vermeiden (Art. 10);
- technische Dokumentation des KI-Systems und seines Zwecks (Art. 11);
- Dokumentation sowie Protokollierung der Vorgänge, insbesondere um die Rückverfolgbarkeit von KI-Ergebnissen zu ermöglichen (Art. 12);
- klare und angemessene Informationen für die Nutzerinnen und Nutzer (Art. 13);
- angemessene menschliche Aufsicht zur Minimierung der Risiken (Art. 14);
- hohes Maß an Robustheit, Sicherheit und Genauigkeit (Art. 15).

Mit geringeren Risiken wird etwa ein Chatbot verbunden, der unter Umständen manipuliert werden könnte. Bei einem solchen System müssen lediglich Transparenzpflichten erfüllt werden, sodass das Verhalten des Chatbots gegebenenfalls nachvollzogen werden kann. Die große Mehrheit der KI-Systeme stellt jedoch nur ein minimales oder gar kein Risiko für die Rechte oder Sicherheit der Bürger dar (z. B. KI-gestützte Videospiele oder Spamfilter). Diese Systeme sind von dem Entwurf der EU-Kommission nicht erfasst und sollen – unter Einhaltung des allgemein geltenden Rechts – weiterhin entwickelt und verwendet werden können.

Digitalisierungsgesetz in Schleswig–Holstein

In Schleswig-Holstein ist seit Frühjahr 2022 das Gesetz über die Möglichkeit des Einsatzes von datengetriebenen Informationstechnologien bei öffentlich-rechtlicher Verwaltungstätigkeit in Kraft. Damit hat Schleswig-Holstein als erstes Bundesland eine explizite Regelung für den Einsatz von KI-Systemen geschaffen. Es werden Anforderungen, darunter Transparenz, Beherrschbarkeit, Robustheit und Sicherheit, sowie Grenzen aufgezeigt. Sogenannte datengetriebene Informationstechnologien sind in bestimmten Anwendungsbereichen nicht zulässig, zum Beispiel zum Zweck der Beurteilungen der Persönlichkeit oder Arbeitsleistung von Menschen und zur

Erstellung von Prognosen über die Straffälligkeit von Personen. Auch bei Ermessen und Beurteilungsspielraum beim Erlass eines Verwaltungsaktes dürfen keine datengetriebenen Informationstechnologien verwendet werden. Das Gesetz unterscheidet zwischen den drei Automationsstufen Assistenzsystem, Delegation und autonome Entscheidung. Eine Zuordnung ist verpflichtend und für die Beurteilung von Risiken sowie geeigneten technischen und organisatorischen Maßnahmen relevant. Weiter sind auch Regelungen zu Transparenz, Menschlicher Aufsicht, Vorrang menschlicher Entscheidungen, der Datengrundlage, der Verarbeitung personenbezogener Daten, Beherrschbarkeit und Risiko, Sicherheit, Robustheit und Resilienz sowie Mindeststandards durch Verordnung enthalten. Darüber hinaus wird die sogenannte KI-Rüge geregelt. Adressaten können dabei innerhalb eines Monats ab Bekanntgabe einer KI-basierten Entscheidung verlangen, dass diese durch eine natürliche Person überprüft wird.

12.3 Die Datenschutz-Grundverordnung

Die Datenschutz-Grundverordnung (DSGVO) findet Anwendung, wenn KI mit personenbezogenen Daten trainiert wird oder diese verarbeitet. In dem Fall gelten die in Artikel 5 DSGVO normierten Grundsätze.

Artikel 5 DSGVO

1. Personenbezogene Daten müssen
   a) auf rechtmäßige Weise, nach Treu und Glauben und in einer für die betroffene Person nachvollziehbaren Weise verarbeitet werden („Rechtmäßigkeit, Verarbeitung nach Treu und Glauben, Transparenz“);
   b) für festgelegte, eindeutige und legitime Zwecke erhoben werden und dürfen nicht in einer mit diesen Zwecken nicht zu vereinbarenden Weise weiterverarbeitet werden; eine Weiterverarbeitung für im öffentlichen Interesse liegende Archivzwecke, für wissenschaftliche oder historische Forschungszwecke oder für statistische Zwecke gilt gemäß Artikel 89 Absatz 1 nicht als unvereinbar mit den ursprünglichen Zwecken („Zweckbindung“);
   c) dem Zweck angemessen und erheblich sowie auf das für die Zwecke der Verarbeitung notwendige Maß beschränkt sein („Datenminimierung“);

d) sachlich richtig und erforderlichenfalls auf dem neuesten Stand sein; es sind alle angemessenen Maßnahmen zu treffen, damit personenbezogene Daten, die im Hinblick auf die Zwecke ihrer Verarbeitung unrichtig sind, unverzüglich gelöscht oder berichtigt werden („Richtigkeit“);
e) in einer Form gespeichert werden, die die Identifizierung der betroffenen Personen nur so lange ermöglicht, wie es für die Zwecke, für die sie verarbeitet werden, erforderlich ist; personenbezogene Daten dürfen länger gespeichert werden, soweit die personenbezogenen Daten vorbehaltlich der Durchführung geeigneter technischer und organisatorischer Maßnahmen, die von dieser Verordnung zum Schutz der Rechte und Freiheiten der betroffenen Person gefordert werden, ausschließlich für im öffentlichen Interesse liegende Archivzwecke oder für wissenschaftliche und historische Forschungszwecke oder für statistische Zwecke gemäß Artikel 89 Absatz 1 verarbeitet werden („Speicherbegrenzung“);
f) in einer Weise verarbeitet werden, die eine angemessene Sicherheit der personenbezogenen Daten gewährleistet, einschließlich Schutz vor unbefugter oder unrechtmäßiger Verarbeitung und vor unbeabsichtigtem Verlust, unbeabsichtigter Zerstörung oder unbeabsichtigter Schädigung durch geeignete technische und organisatorische Maßnahmen („Integrität und Vertraulichkeit“);

2. Der Verantwortliche ist für die Einhaltung des Absatzes 1 verantwortlich und muss dessen Einhaltung nachweisen können („Rechenschaftspflicht“).

Die DSGVO enthält keine spezifischen Pflichten für KI-Systeme, zur Einhaltung der Verordnung und deren Grundsätze ist die öffentliche Verwaltung insgesamt verpflichtet, unabhängig von der Nutzung künstlicher Intelligenz. Verstöße gegen die Grundsätze des Artikel 5 können ein Bußgeld nach sich ziehen (Art. 83 Abs. 5 lit. A DSGVO). Bei automatisierten Entscheidungen sieht die DSGVO weitreichendere Maßnahmen zur Wahrung der Rechte, Freiheiten und Interessen der Betroffenen vor. Diese müssen darüber informiert werden, dass eine automatisierte Einzelentscheidung getroffen wurde und welche involvierte Logik hinter der Entscheidung steckt. Die Informationen sind in präziser, transparenter, verständlicher und leicht zugänglicher Form in einer klaren und einfachen Sprache der betroffenen Person kostenfrei zur Verfügung zu stellen.

In Artikel 4, Nr. 1 definiert die DSGVO personenbezogene Daten als

„alle Informationen, die sich auf eine identifizierte oder identifizierbare natürliche Person (im Folgenden „betroffene Person“) beziehen; als identifizierbar wird eine

natürliche Person angesehen, die direkt oder indirekt, insbesondere mittels Zuordnung zu einer Kennung wie einem Namen, zu einer Kennnummer, zu Standortdaten, zu einer Online-Kennung oder zu einem oder mehreren besonderen Merkmalen, die Ausdruck der physischen, physiologischen, genetischen, psychischen, wirtschaftlichen, kulturellen oder sozialen Identität dieser natürlichen Person sind, identifiziert werden kann“.

Als personenbezogene Daten sind unter anderem anerkannt: Namen und Vornamen natürlicher Personen sowie deren Einkommen, deren Geburtsdatum, Geschlecht, ethnische Zugehörigkeit, Religion und Sprache, deren Staatsangehörigkeit, Melde- und Aufenthaltsstatus sowie Passangaben, Aufzeichnungen über die Arbeitszeit oder Fingerabdrücke. Um personenbezogene Daten zu anonymisieren, muss der Personenbezug derart aufgehoben werden, dass eine erneute Re-Identifizierung nicht oder nur mit einem unverhältnismäßig hohen Aufwand wieder hergestellt werden kann (Brink & Wolff, 2021, Rn. 15a).

Vorselektierung von Beweismaterial

Seit Anfang des letzten Jahres setzt die Polizei in Niedersachsen eine Software auf Basis von Künstlicher Intelligenz ein, die durch IT-Experten des LKA Niedersachsen entwickelt und darauf trainiert wurde, relevante von irrelevanten Daten zu trennen (Landeskriminalamt Niedersachsen, 2020). Die Software „Niki“ unterstützt damit Polizeibeamte im Kampf gegen Kinderpornographie, indem sie dabei hilft, Beweismaterial auszuwerten. Sie kommt damit erst dann zum Einsatz, wenn Bild- und Videomaterial bei einem Tatverdächtigen sichergestellt oder beschlagnahmt wurde. Im Rahmen der Ermittlungsarbeit werden Polizeibeamte oft mit Datenmengen konfrontiert, die gerade in den letzten Jahren ein kaum mehr handhabbares Niveau erreicht haben. Die polizeiliche Erfahrung zeigt, dass in der Regel bis zu 75 % der sichergestellten Bilder nicht relevant sind. Ein Viertel des Bildmaterials ist aber regelmäßig strafrechtlich relevant und wenn man diese Bilder zeitnah identifizieren kann, können die Ermittlungen erfolgreicher und schneller werden (Strünkelnberg & Michel, 2020). Dadurch ist es möglich, noch andauernden Missbrauch schnell zu beenden. Sobald das Programm eine Vorauswahl getroffen hat, müssen Polizeibeamte die Detailauswertung übernehmen. Die ersten Rückmeldungen aus den Flächenbehörden in Niedersachsen sind positiv. Nach Angaben der Behörden wird die Software den Erwartungen gerecht und „nicht relevantes“ Bildmaterial wird erfolgreich vorselektiert. Die Software wurde im Rahmen des Programms „Polizei 2020“ inzwischen auch anderen Polizeibehörden aus Bund und Ländern zur Verfügung gestellt.

Bei der Vorselektierung von Bildern zur Identifizierung von kinderpornographischem Inhalt werden zwar personenbezogene Daten gesichtet, da die Dateien auf einem zugehörigen Rechner einer Person abgespeichert sind, eine Sichtung durch die Polizeibehörde ist dennoch erlaubt, da diese Ermittlungstätigkeit nicht in den sachlichen Anwendungsbereich der DSGVO fällt. Gemäß Artikel 2 Absatz 2d findet die Verordnung keine Anwendung auf die Verarbeitung personenbezogener Daten, soweit die zuständige Behörde zum Zwecke der Verhütung, Ermittlung, Aufdeckung oder Verfolgung von Straftaten tätig wird.◀

12.4 Übung zur Datenschutz-Grundverordnung

Manchmal ist es auf den ersten Blick gar nicht eindeutig erkennbar, dass es sich um personenbezogene Daten handelt. Handelt es sich bei diesen Beispielen um personenbezogene Daten oder nicht (vgl. Brink & Wolff, 2021)?

- Tags bei Graffitis
- Die Email-Adresse auf der Website eines Unternehmens in der Form info@unternehmen.de
- GPS-Daten zur Standortermittlung von Firmenfahrzeugen
- Die durchschnittliche Anzahl der Besucher eines Wochenmarktes
- Dynamische IP-Adressen

12.5 Die Vereinbarkeit von KI und Datenschutz am Beispiel eines Chatbots

Bei der Implementierung und dem Betrieb von KI-Systemen fehlen bisher noch weitgehend einheitliche Leitlinien, weshalb es häufig zu extremen Positionen rund um den Einsatz von KI kommt. Während einige diesen technischen Fortschritt mit „zu gefährlich“ abtun, sehen andere KI als „Lösung aller Probleme“. Wichtig ist es, zu einer realistischen Einschätzung von Risiken zu gelangen und dann entsprechende Gegenmaßnahmen zu ergreifen. Hierfür ist ein Wissen über die rechtlichen Vorgaben zum Einsatz von KI notwendig. Beim unterstützenden Einsatz von KI, wenn es also nicht um fachliche Entscheidungen geht – wie etwa im

Rahmen der Bürgerkommunikation bei einfachen Anfragen mittels eines Chatbots –, beschränken sich die Risiken weitgehend auf die allgemeinen Risiken, die grundsätzlich mit dem Einsatz von IT einhergehen.

Die wichtigste Rechtsvorschrift in diesem Bereich ist daher die Datenschutz-Grundverordnung (DSGVO). Die DSGVO ist unmittelbar geltendes Recht und hat Vorrang vor nationalen Regelungen. Das Bundesdatenschutzgesetz ergänzt daher die DSGVO nur um die Bereiche, in denen die EU-Verordnung den Mitgliedstaaten Gestaltungsspielräume belässt. Die DSGVO enthält keine KI-spezifischen Regelungen. Sie findet jedoch immer dann Anwendung, wenn KI-Systeme personenbezogene Daten verwenden. Chatbots bewegen sich fast immer im Anwendungsbereich der DSGVO, da Gesprächsdaten und Kommunikationsdaten (wie etwa die IP-Adresse des Computers) personenbezogene Daten im Sinne der DSGVO sind. Daher gelten auch beim Einsatz von Chatbots die in Artikel 5 DSGVO normierten Grundsätze für die Verarbeitung personenbezogener Daten. Der Begriff Grundsätze ist insoweit etwas unglücklich gewählt. Es handelt sich nämlich keineswegs nur um grobe Leitlinien, sondern um verbindliche Regelungen. Es ist daher treffender, von Grundpflichten zu sprechen, deren Verwirklichung verpflichtend ist und nicht nur bestmöglich angestrebt werden muss.

Zu den wichtigsten Grundpflichten gehört, dass personenbezogene Daten nur auf rechtmäßige Weise verarbeitet werden dürfen. Das bedeutet, dass vor der Nutzung des Chatbots die Einwilligung der betroffenen Person zur Erhebung und Nutzung seiner oder ihrer personenbezogenen Daten eingeholt werden muss. Die Daten dürfen zudem nur für den festgelegten Zweck verarbeitet werden. Eine Verarbeitung zu noch unbekannten Zwecken scheidet daher aus. Außerdem ist der Grundsatz der Datenminimierung einzuhalten. Angaben, die für die jeweilige Funktion des Chatbots nicht erforderlich sind, dürfen auch nicht erhoben werden. Die namentliche Ansprache kann zwar ein Gefühl des persönlichen Gesprächs vermitteln, doch eine Ansprache mit dem Namen der Nutzerinnen und Nutzer ist gerade bei einfachen Anfragen, z. B. nach Öffnungszeiten, häufig nicht notwendig. Daneben müssen die Daten richtig sein, also die Realität zutreffend abbilden. Dieser Aspekt ist weniger wichtig, wenn es allein um die reine Kommunikation via Chatbot geht. Sie ist aber umso wichtiger, wenn es um die Informationsgewinnung für eine anschließende Entscheidung geht. Die betroffene Person darf durch die Verwendung fehlerhafter Daten keine Nachteile erleiden. Des Weiteren muss eine „Datensparsamkeit“ auch in zeitlicher Hinsicht gewährleistet sein. Das bedeutet, dass die Daten nach Zweckerreichung wieder gelöscht werden müssen. Personenbezogene Daten sind zudem so zu verarbeiten, dass eine angemessene Sicherheit dieser Daten gewährleistet ist. Hierzu hat die

verantwortliche Person geeignete technische und organisatorische Maßnahmen zu ergreifen, um den Schutz vor unbefugter oder unrechtmäßiger Verarbeitung und vor unbeabsichtigtem Verlust zu gewährleisten.

Wie bei jedem datenschutzrechtlich relevanten Vorgang, müssen auch im Rahmen der Chatbot-Nutzung die Informationspflichten aus Artikel 13 und 14 DSGVO eingehalten werden. Es müssen daher geeignete Maßnahmen getroffen werden, um der betroffenen Person die gesetzlichen Pflichtangaben in präziser und leicht zugänglicher Form in klarer und einfacher Sprache zu übermitteln. Dazu gehört grundsätzlich auch die Vermittlung der groben Funktionsweise des eingesetzten KI-Systems. Da die Datenschutzerklärung jederzeit erreichbar eingebunden werden muss, drohen die Vorteile einer unkomplizierten und damit schnellen Kommunikation abhanden zu kommen. Um das zu verhindern, empfiehlt es sich nur Kurzinformationen in den Chatbot einzubinden und im Übrigen auf ergänzende Hinweise zu verlinken.

Die DSGVO versucht durch datenschutzrechtliche Rahmenbedingungen die Interessen der Nutzerinnen und Nutzer miteinander in Einklang zu bringen. Entwicklung und Einsatz von KI geraten jedoch mit nahezu allen in der DSGVO niedergelegten Grundsätzen der Datenverarbeitung in Konflikt. Daher bleibt abzuwarten, ob der von der EU-Kommission geplante europäische Rechtsrahmen für KI auch eine Änderung der Datenschutz-Grundverordnung beinhalten wird. Bis dahin müssen aber alle datenschutzrechtlichen Maßnahmen im Vorfeld strikt umgesetzt werden, auch wenn dies zunächst zu Mehrarbeit führt. Nur so können eine erhöhte Sicherheit gewährleistet und die datenschutzrechtlichen Sanktionen vermieden werden.

12.6 Der vollständig automatisierte Verwaltungsakt

Der unterstützende, „entscheidungsferne“ Einsatz von KI-Systemen in der Verwaltung – etwa die Bürgerkommunikation oder das Übertragen von Daten in Papierform in ein digitales Format – sind rechtlich weniger bedenklich, weil die finale Entscheidung beim Menschen bleibt. Die Risiken beschränken sich hierbei weitgehend auf die allgemeinen Risiken, die mit einem Einsatz von IT-Systemen verbunden werden.

Anders verhält es sich allerdings bei einem Einsatz, der mit nach außen wirksamen Entscheidungen verbunden ist. In diesem Fall ist die rechtliche Bewertung komplexer, weil der Einsatz mit umfangreichen Risiken verbunden ist. Außerdem ist eine Ermächtigungsgrundlage erforderlich. Eine besondere Dimension

wird erreicht, wenn es sich darüber hinaus um einen vollständig automatisierten Verwaltungsakt (VA) handelt. Voraussetzung hierfür ist das Fehlen einer personellen Bearbeitung bei allen Verfahrensschritten innerhalb der Verwaltung. Von einem vollständig automatisierten Erlass geht das Gesetz also nur dann aus, wenn die Entscheidung wirklich vollständig in die Hände des KI-Systems gelegt wird. Der Erlass eines Ermessensverwaltungsaktes mithilfe automatischer Einrichtungen ist daher möglich und auch gewünscht. Mit § 35a Verwaltungsverfahrensgesetz (VwVfG) hat der Gesetzgeber den vollautomatisierten Erlass eines Verwaltungsakts ermöglicht – allerdings nur unter der Bedingung, dass dies durch Rechtsvorschrift zugelassen ist und weder ein Ermessen noch ein Beurteilungsspielraum besteht (vgl. Schoch & Schneider, 2020).

§ 35a Vollständig automatisierter Erlass eines Verwaltungsaktes

Ein Verwaltungsakt kann vollständig durch automatische Einrichtungen erlassen werden, sofern dies durch Rechtsvorschrift zugelassen ist und weder ein Ermessen noch ein Beurteilungsspielraum besteht.

Der Gesetzgeber versucht mit dieser Regelung den technischen Fortschritt und die das Verwaltungsrecht prägenden Prinzipien in Einklang zu bringen. Die Vorschrift zeigt, dass der Gesetzgeber die Verwendung moderner Informationstechnik in Verwaltungsverfahren grundsätzlich begrüßt und daher erleichtern möchte – aber nur in den genannten Grenzen. Und diese Grenze ist der vollständig automatisierte Erlass eines Verwaltungsaktes mit Ermessens- oder Beurteilungsspielraum. Das bedeutet, dass unabhängig von den technischen Möglichkeiten ein vollautomatisierter Erlass eines Verwaltungsaktes mit Ermessens- oder Beurteilungsspielraum nicht möglich sein wird. Auf Gebieten wie dem Polizeirecht, dessen Ermächtigungsgrundlagen den Behörden fast durchgängig Ermessen einräumen, scheidet eine Vollautomatisierung damit auch in Zukunft aus.

Ein vollständig automatisierter Erlass eines Verwaltungsaktes kommt folglich nur bei gebundenen Entscheidungen infrage (vgl. Martini & Nink, 2017). Nur bei diesen ist ein vollständiger Verzicht auf eine personelle Bearbeitung vertretbar, weil eine eindeutige Zuordnung von Rechtsfolgen zu dem festgestellten Sachverhalt möglich ist. Sowohl die Ausübung eines Ermessens als auch die Ausfüllung eines Beurteilungsspielraums erfordern dagegen eine individuelle Abwägung und Willensbetätigung, die ein KI-System – zumindest derzeit – nicht leisten kann.

Was sich dem Gesetzeswortlaut nicht eindeutig entnehmen lässt, ist die Frage, was unter dem Begriff „automatische Einrichtung“ genau zu verstehen ist. Einigkeit besteht insoweit, dass der Begriff technikoffen so formuliert ist, dass letztlich die Verwendung jeglicher technischer Einrichtungen umfasst ist, die nach vorher festgesetzten Parametern ohne weiteres menschliches Einwirken funktionieren. Eine eindeutige Beschränkung auf regelbasierte Expertensysteme lässt sich weder der Norm selbst noch der Gesetzesbegründung entnehmen. Es ist daher davon auszugehen, dass auch selbstlernende Algorithmen unter diesen Begriff fallen. Entgegenstehende Rechtsprechung gibt es bislang nicht.

> § 24 Untersuchungsgrundsatz (1) […] Setzt die Behörde automatische Einrichtungen zum Erlass von Verwaltungsakten ein, muss sie für den Einzelfall bedeutsame tatsächliche Angaben des Beteiligten berücksichtigen, die im automatischen Verfahren nicht ermittelt würden.

Das Gesetz verlangt an dieser Stelle, dass bedeutende tatsächliche Angaben der Beteiligten „händisch“ berücksichtigt werden müssen, wenn eine Ermittlung im automatischen Verfahren nicht möglich ist. Mit der Automatisierung geht zwangsläufig eine Schematisierung einher. Die jeweiligen individuellen Komponenten einer Fallkonstellation können daher bei einer automatisierten Entscheidung nur einbezogen werden, wenn sie zuvor ermittelt und bei der Ausgestaltung der automatischen Einrichtung berücksichtigt werden (vgl. auch Guckelberger, 2021). In der Praxis bedeutet dies, dass der vollständig automatisch generierte Verwaltungsakt bei Bedarf nachträglich abgeändert werden kann, oder eine weitere Bearbeitung außerhalb des automatisierten Verfahrens möglich sein muss. Eine Bearbeitung außerhalb des automatisierten Verfahrens ist aber nur dann erforderlich, wenn die Angaben für den Einzelfall tatsächlich von Bedeutung sind.

Für eine effektive Ausübung des Rechts auf Anhörung gilt auch bei vollautomatisierten Verwaltungsverfahren das Recht der Akteneinsicht gemäß § 29 VwVfG. Dies ist zugegebenermaßen nicht leicht zu verwirklichen. Bereits bei der Konzipierung der automatischen Einrichtungen muss dafür gesorgt werden, dass die wesentlichen Abläufe dokumentiert werden, die zur getroffenen Entscheidung geführt haben. Besondere Schwierigkeiten ergeben sich bei selbstlernenden Algorithmen, bei denen der Weg der Entscheidungsfindung zum Teil nur schwer nachvollzogen werden kann. Hier zeichnet sich in Zukunft für den Fachgesetzgeber erheblicher Regelungsbedarf im Zusammenhang mit der Einführung solcher

Systeme ab. Denkbar sind zum Beispiel fachrechtliche Vorgaben für stichprobenartige, zu protokollierende Kontrollen oder die Verpflichtung zur Offenlegung der maßgeblichen Programmcodes sowie der genutzten Daten.

§ 35a VwVfG gilt nicht nur für den Erlass von Verwaltungsakten im Ausgangsverfahren, sondern grundsätzlich auch für Änderung, Rücknahme und Widerruf. Wird ein vollautomatisiertes Verwaltungsverfahren unter Nichtbeachtung des Rechtsvorschriftenvorbehalts und sonstiger Grenzen eingeführt, sind die so erlassenen Verwaltungsakte allein deshalb rechtswidrig und mithin vom Bürger angreifbar.

12.7 Übung zum vollständig automatisierten Verwaltungsakt

1. Sie möchten einen vollständig automatisierten Verwaltungsakt für einen bestimmten Verwaltungsprozess implementieren. Worauf müssen Sie achten?
   a) § 35a VwVfG stellt eine Ermächtigungsgrundlage für den Erlass vollautomatisierter Verwaltungsakte dar, deshalb muss rechtlich nichts weiter beachtet werden.
   b) Automatisierte Verwaltungsakte sind nur für ein Ausgangsverfahren möglich, nicht aber für Änderung, Rücknahme oder Widerruf.
   c) Automatisierte Verfahren müssen zuvor durch „Rechtsvorschrift“ (formelle Gesetze, Rechtsverordnungen oder Satzungen) zugelassen werden.
   d) Beim Umgang mit personenbezogenen Daten muss immer die DS-GVO berücksichtigt werden.
2. Artikel 5 DS-GVO enthält spezifische Vorschriften für KI-Systeme.
   a) Diese Aussage ist korrekt.
   b) Diese Aussage ist falsch.
3. Welche Aussage zum Datenschutz ist zutreffend?
   a) Datenschutz ist für den Einsatz von KI-Systemen nicht relevant.
   b) Verstöße gegen die Grundsätze des Artikel 5 können ein Bußgeld nach sich ziehen (Art. 83 Abs. 5 lit. A DS-GVO).
   c) Wenn personenbezogene Daten verwendet werden, gelten die Informationspflichten aus Artikel 13 und 14 DS -GVO.
   d) Chatbots bewegen sich fast immer im Anwendungsbereich der DS-GVO, da Gesprächsdaten und Kommunikationsdaten (wie etwa die IP-Adresse des Computers) personenbezogene Daten im Sinne der DS -GVO sind.

4. Welche Aussagen über den Verordnungsentwurf der EU-Kommission sind zutreffend?
   a) Der Entwurf hat als Grundlage ein zuvor von der Kommission erstelltes Weißbuch genutzt.
   b) Im Entwurf wird unterschieden, ob mit der Nutzung eines KI-Systems ein hohes, geringes oder minimales Risiko verbunden ist.
   c) Im Entwurf werden Szenarien beschrieben, für die ein KI-Einsatz verboten wird.
   d) Der Entwurf ist niedergeschriebenes Völkergewohnheitsrecht.
5. Herr Müller arbeitet seit kurzer Zeit in der Steuerverwaltung. Er weiß, dass die Steuerverwaltung häufig als Vorreiter innerhalb der deutschen Verwaltung im Kontext von Digitalisierung und Automatisierung bezeichnet wird und das Steuerangelegenheiten regelmäßig vollständig elektronisch abgewickelt werden. Herr Müller ist sich aber ganz sicher, dass er damals gelernt hat, dass das Verwaltungsverfahrensgesetz – und damit § 35a – nicht für die Abgabenordnung gilt. Aber wie kann es dann sein, dass die Steuerverwaltung vollständig automatisierte Entscheidungen treffen kann?
   a) Herr Müller erinnert sich richtig, da gibt es diese Vorschrift im VwVfG die besagt, dass dieses Gesetz keine Anwendung auf die Abgabenordnung findet. Allerdings hat der Gesetzgeber zusammen mit § 35a VwVfG auch den § 155 Abs. 4 AO (Abgabenordnung) eingeführt. Der ist zwar ein bisschen anders formuliert, ermöglicht aber auch den Erlass eines vollständig automatisierten Verwaltungsakts.
   b) Herr Müller täuscht sich, dass Verwaltungsverfahrensgesetz findet auch auf die Abgabenordnung Anwendung. Schließlich ist das ja auch öffentliches Recht…

12.8 Aufgaben zum eigenen Anwendungsfall

- Wie stellen Sie sicher, dass die von Ihnen genutzten Daten konform sind mit der Datenschutz-Grundverordnung?
- Der Entwurf der EU-Kommission zur Reglementierung von KI-Systemen unterteilt diese in unterschiedliche Risikogruppen. In welche Risikogruppe fällt Ihr System? Erläutern Sie Ihre Einschätzung.

Literatur

Brink, S., & Wolff, H. A. (o. J.). *BeckOK Datenschutzrecht* (37. Edition, Stand: 01.08.2021).

Europäische Kommission. (Hrsg.). (2020). *Weißbuch Zur Künstlichen Intelligenz – ein europäisches Konzept für Exzellenz und Vertrauen* (COM (2020) 65 final). https://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_de.pdf. Zugegriffen: 15. Okt. 2022.

Guckelberger, A. (2021). Automatisierte Verwaltungsentscheidungen: Stand und Perspektiven. In *Die Öffentliche Verwaltung: Bd. Die Öffentliche Verwaltung* (S. 566–578). Kohlhammer.

Hähnchen, S., Schrader, P. T., Weiler, F., & Wischmeyer, T. (2020). Legal Tech. *Jus: Juristische Schulung, 7,* 625–635.

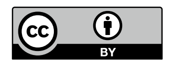
Landeskriminalamt Niedersachsen. (Hrsg.). (2020). *Künstliche Intelligenz: LKA Niedersachsen stellt Software zur Bekämpfung von Kinderpornografie bundesweit zur Verfügung* (Presseinformation Nr. 6) [Pressemitteilung]. Landes Kriminalamt Niedersachsen. https://www.lka.polizei-nds.de/a/presse/pressemeldungen/kuenstliche-intelligenz-lka-niedersachsen-stellt-software-zur-bekaempfung-von-kinderpornografie-bundesweit-zur-verfuegung-114750.html. Zugegriffen: 15. Okt. 2022.

Martini, M., & Nink, D. (2017). Wenn Maschinen entscheiden... – Vollautomatisierte Verwaltungsverfahren und der Persönlichkeitsschutz. *Neue Zeitschrift für Verwaltungsrecht – Extra, 36*(10), 1–14.

Schoch, F., & Schneider, J.-P. (2020). Verwaltungsrecht VwVfG Band III Kommentar. C.H. Beck-Verlag. Die Kommentierung des § 35a VwVfG ist von Prof. Dr. Gerrit Hornung erfolgt.

Strünkelnberg, T., & Michel, R. (16. Januar 2020). Niedersachsen setzt im Kampf gegen Kinderpornografie auf KI. Weser Kurier. https://www.weser-kurier.de/bremen/niedersachsen-setzt-im-kampf-gegen-kinderpornografie-auf-ki-doc7e3kbp6tcm09aonhjmu. Zugegriffen: 15. Okt. 2022.

Vorschlag für eine Verordnung des Europäischen Parlaments und des Rates zur Festlegung harmonisierter Vorschriften für Künstliche Intelligenz (Gesetz über Künstliche Intelligenz) und zur Änderung bestimmter Rechtsakte der Union, 21.04.2021. https://eur-lex.europa.eu/legal-content/DE/TXT/HTML/?uri=CELEX:52021PC0206&from=EN.


BY



Ausblick 13

Zusammenfassung

Mit Blick auf die Nutzung KI-basierter Lösungen befinden wir uns derzeit in einer frühen Phase mit wenigen Pilotanwendungen. Eine besondere Herausforderung liegt in der Identifikation geeigneter Anwendungsfelder. Diese Technologieorientierung ist in der Erprobungsphase nachvollziehbar, geht jedoch häufig an den Problemen und Anforderungen der Verwaltungspraxis vorbei. Es besteht die Gefahr, dass alternative Lösungsansätze vernachlässigt werden.

Initiativen zur Identifikation geeigneter Anwendungsfelder in Forschung, Wirtschaft und im öffentlichen Sektor sind auf eine enge Zusammenarbeit zwischen KI-Experten und Fachexperten aufseiten der Verwaltungen angewiesen. Probleme der Verwaltungspraxis sollten die Ausgangslage bilden, anhand derer geprüft wird, welche Lösungsszenarien – darunter auch KI-basierte Anwendungen – denkbar sind. Dabei sind verschiedene Fragen zu beantworten:

- Welche Ziele werden mit dem KI-Einsatz verbunden?
- Welche technischen Voraussetzungen müssen erfüllt werden?
- Liegen die notwendigen Daten in ausreichender Qualität vor?
- Wie sollte das KI-System in die bestehenden Prozesse integriert werden?
- Welche Grenzen werden durch den rechtlichen Rahmen gesetzt?
- Welche organisatorischen Änderungen gehen mit der Lösung einher?
- Welche neuen Anforderungen ergeben sich gegenüber Betroffenen und Beschäftigten?

M. Heine et al., *Künstliche Intelligenz in öffentlichen Verwaltungen*, Edition eGov-Campus, https://doi.org/10.1007/978-3-658-40101-6_13

- Welche Wirkung entfaltet der Einsatz von KI-Anwendungen (auf Ebene der Betroffenen, der Beschäftigten und auf gesellschaftlicher Ebene)?
- Wer ist für die Planung, Entwicklung bzw. Beschaffung und den Betrieb zuständig?
- Ist der Einsatz wirtschaftlich und gemeinwohlorientiert?

Insbesondere mit Blick auf die Zielstellung, die mit dem KI-Einsatz verbunden ist, muss eine Evaluation im Pilotbetrieb bzw. im Betrieb durchgeführt werden – auch um die Wirkung des Systems bewerten zu können. Der Transfer bereits bewährter Lösungen ist ein wesentlicher Schlüssel zu einem angemessenen Fortschritt bei der Anwendung von KI-Lösungen im öffentlichen Sektor.

Die Fragen zur Identifikation geeigneter Anwendungsszenarien verdeutlichen abermals, dass eine enge Zusammenarbeit zwischen KI-Experten und Fachexperten erforderlich ist. Dies setzt ein gemeinsames Grundverständnis voraus. Investitionen in Aus- und Weiterbildung sind daher ein zentraler Schritt, der derzeit in vielen Verwaltungen angegangen wird. Dabei darf KI in öffentlichen Verwaltungen nicht isoliert betrachten werden. Digitalisierung und auch der Einsatz von KI-Anwendungen dienen der Verwaltungsreform mit dem Ziel, Verbesserungen in den Dimensionen Zeit, Kosten und Qualität zu erreichen.

